Quantifying Intestinal Stem Cell Dynamics Using Microsatellite Sequencing

Joseph A. Christopher

Cancer Research UK Cambridge Institute

Queens' College
University of Cambridge

This dissertation is submitted for the degree of

Doctor of Philosophy

August 2016

Mum, Dad and Holly.
For you. Because of you.

Declaration

I hereby declare that except where specific reference is made to the work of others, the contents of this dissertation are original and have not been submitted in whole or in part for consideration for any other degree or qualification in this, or any other university. This dissertation is my own work and contains nothing which is the outcome of work done in collaboration with others, except as specified in the text and Acknowledgements. This dissertation contains fewer than 60,000 words including appendices, bibliography, footnotes, tables and equations and has fewer than 150 figures.

Joseph A. Christopher
August 2016

Acknowledgements

Doug, your patience, support and guidance have been invaluable throughout my time in your group. Thank you for making the past 4 years such an enjoyable and enlightening experience. You have fuelled my passion for research and I will always be grateful for that.

Richard, your mastery of molecular biology and eye for detail have saved me countless hours through inspired insight and experimental design advice. Thank you for spending even more countless hours going through this dissertation with a fine-tooth comb.

Ed and Lee, your hours of teaching and guidance in the computational aspects of this project have enabled me to work independently and, as a result, progress this project far further than it would have done otherwise.

Sarah, your detailed record taking and tolerance of countless, often obscure, questions relating to the $[CA]_{30}$ reporter project have, very literally, made this project possible. Your support and concern have kept me calm throughout.

Filipe, thank you for all your help with the $Msh2^{fl/fl}$ mice. I have thoroughly enjoyed our collaborative efforts in the lab.

Sofie, thank you for cloning and screening dozens, if not hundreds, of plasmids containing $[CA]_n$ alleles. You saved me many hours of laborious work.

To the entire Winton group, thank you for your continued support and friendship. You have shared in my lows and celebrated my highs. Working with you has been a true highlight of my time in Cambridge.

To all of the staff and students at Cancer Research UK Cambridge Institute (CRUK-CI), it has been an absolute pleasure to share in your community. I have made friends for life and thrived in the environment created by you all.

And last but by no means least, thank you to all of the CRUK-CI core facilities, management and services. You are absolutely the reason for the Institute's continued success. In particular, I would like to thank Genomics for their technical advice and sequencing of dozens of libraries; James Hadfield for his help and advice designing the custom M13

sequencing adaptor set; Bioinformatics and IT for keeping the cluster and demultiplexing pipelines up to date and running smoothly; Research Instrumentation for regularly sharing their extensive technical know-how with me; all of Lab Management for facilitating my work on a day to day basis and the Biological Resource Unit (BRU) for ensuring the highest standards of animal welfare for all of the mice used in my study.

Abstract

The intestinal epithelium is rapidly renewing throughout life. A population of stem cells exist within the intestinal crypt that drive rapid cell renewal and replace each other by a pattern of neutral drift. Perturbation of these dynamics through oncogenic mutation can predispose the epithelium to neoplastic transformation. Understanding the factors that govern these dynamics will give insight into the early stages of oncogenesis.

Continuous clonal labelling, whereby DNA strand slippage leading to the contraction or expansion of a microsatellite during mitotic replication, can be employed to enable the detection of a single clone. Previous studies have shown that quantification of clone size over time allows inference of the functional stem cell number and stem cell replacement rate within intestinal crypts.

Current continuous labelling techniques require the introduction of a transgenic microsatellite into a model system genome, such as mouse, that leads to reporter expression following mutation. This obviously precludes human studies. Alternative somatic alterations techniques used for continuous labelling in humans requires spontaneous loss of a protein, or change in methylation status, within a single clone that can act as a clonal mark. Though these techniques have given insight into the spread of mutations within the intestinal epithelium and enabled inference of adenoma clonality, the true neutrality of these changes are currently unknown. We propose that the small changes in endogenous microsatellite length will act as a neutral clonal mark within the intestinal epithelium and allow an unbiased approach to quantifying intestinal stem cell dynamics in human intestinal tissues.

To overcome the many technical challenges associated with accurate measurement of microsatellite length, a stepwise approach was taken to develop a technique for the multiplexed high throughput sequencing of up to 21 native dinucleotide repeats in hundreds of single crypts. Furthermore, a novel method was developed for the quantification of clone size from data generated from the targeted re-sequencing of microsatellites in single crypts.

This protocol was validated *in vitro* and *in vivo* in mouse. Furthermore, proof of prin-

ciple sequencing in human crypts was performed to show that this method is suitable for larger scale quantification of intra-cryptal clone size in human tissue.

This, and similar approaches, may be the only way to quantify intestinal stem cell dynamics within the healthy human colon or, dysplastic or adenomatous patient tissue. These measurements should give a unique insight into the dynamics of healthy, pre-neoplastic and neoplastic human intestinal stem cells.

Table of contents

List of figures

List of tables

Chapter 1

Introduction

The early pioneers of modern cellular biology placed the *cellula* as the basic building block of an organism, a central tenant of 'cell theory'. Subsequent observations of binary fission of the cell and nucleus, particularly in injured tissues, lead to the formation of the first theories regarding regeneration of tissue via cell division pioneered by Walther Flemming [70]. These theories were bolstered by the work of Giulio Bizzozero in the late 19^{th} century with experiments showing an increase in 'a blood-cell-forming phenomenon in the bone marrow' following controlled bleeding experiments in chickens and pigeons. Bizzozero went onto catalogue many tissues of the body using the frequency of mitotic bodies as an indicator of a tissue's regenerative capacity. He categorised all tissues into three different types: 'labile' to describe those undergoing continuous regeneration, 'stable' to describe those undergoing division up until birth and 'everlasting' where no cell division was observed [62]. Even in these early observations, it was very clear that the gut was extraordinarily proliferative and so was categorised as 'labile'. However, the question still remained as to the origin of these newly differentiated cells.

Early work by Ernst Haeckel in 1868 coined the term 'stammzelle' or 'stem cell' and put forward the first biological definition of a stem cell being a cell that can give rise to all cells of a multicellular organism. This work was built upon by August Weissman in the late 19^{th} century in which he proposed his theory of the continuity of germ-plasm. Weissman proposed that early in embryonic development there was segregation that lead to the development of specialised germ cells that were distinct from the the rest of the somatic cells of the body. This definition of a stem cell was developed further by Boveri in 1892 to include non-committed cells between the zygote and the germ cells. It would seem that these early stem cell biologists were actually describing what we call primordial germ cells and

germline stem cells today [80]. The existence of the adult stem cell was not fully appreciated until decades later.

The concept of an adult stem cell was first applied by those studying the haematopoietic system in the early 20th century. As early as 1957, the pioneering work of E. Donnall Thomas and others had brought bone marrow transplant to the clinic long before the haematopoietic stem cell had been fully described [98]. The first definitive evidence of a haematopoietic cell population with multipotent properties was not provided until the experiments performed by Till and McCulloch in the early 1960s [5, 100, 101]. Inspired by this work, biologists began the search for the adult stem cell of other tissues. Among the first to be described outside the haematopoietic system were the muscle-derived satellite cells [61, 63], bone marrow-derived mesenchymal stem cells [35] and intestinal stem cells [24].

Seminal experiments with ^{3}H-thymidine treatment of mice by Cheng and Leblond in the 1970s provided strong evidence for the crypt-base columnar cells (CBCCs) as the adult stem cell of the small intestine and colon. Through use of labelled phagosomes within the CBCC population, Cheng and Leblond were able to lineage trace from this population and show that all four main epithelial cell types originated from this population. Furthermore, these experiments enabled quantification of cell turnover within the gut epithelium, which confirmed Bizzozero's earlier observations that the gut epithelium has one of the highest cell turnovers of any adult mammalian tissue with the vast majority of epithelial cells being lost every 2-4 days [20–24]. This work clearly showed a common stem cell for most epithelial lineages but it was not until the early 1990s that the presence of a separate neuroendocrine progenitor was disproved [99] and a common stem cell for all intestinal epithelium lineages was postulated. This work largely put to rest any theory pertaining to extrinsic cellular contribution to cell turnover within the adult gut epithelium and paved the way for the molecular and functional characterisation of the adult intestinal stem cell.

1.1 The intestinal epithelium

The intestinal epithelium is arranged in a monolayer consisting of multiple differentiated cell types that together enable the co-ordinated digestion and absorption of ingested food and the concomitant excretion of waste products. Juxtaposed to the intestinal epithelium is supporting stroma consisting of myofibroblasts, immune cells, lymph vessels and vasculature. Together these cell types make up the core cellular components of the intestinal

tract and their continued function are essential for mammalian life as we know it. Throughout the intestinal epithelium there are regular invaginations that create structures known as the crypts of Leiburkühn. The small intestine is well adapted to the absorption of nutrients through massive increase in available surface area generated by folds of epithelium that protrude into the intestinal lumen known as villi. The colon, in contrast, has a flat surface epithelium more suited to smooth transit of waste products and water absorption. The overall structure and organisation of the gut epithelium is well adapted to sustain nutrient digestion and absorption whilst maintaining remarkable proliferative and regenerative capacity.

The intestinal epithelium consists of four major cell types. The majority of cells are enterocytes which are specialised for the absorption of nutrients as well as the expression of key digestive enzymes on their apical surfaces. Within the crypt base, in the small intestine only, are a population of Paneth cells that secrete bactericidal defensin peptides and lysozymes that protect the gut epithelium from pathogenic bacteria. Additionally, the Paneth cells provide a source of Wnt ligands that support the multipotent intestinal stem cell population within the crypt base [85]. In the colon, Paneth cells are not found and instead colonic Crypt Base Secretory Cells are postulated to be Paneth-like cells that constitute the stem cell niche [83]. The third major cell type is the goblet cell identified by mucous granules that accumulate at the apical membrane for secretion. Secretion of mucous by these cells protects the duodenum from acid damage and later aids the passage of stool through the colon. The final major cell type is the enteroendocrine cell which is identifiable by a basal build-up of secretory vesicles that co-ordinate systemic and local regulation of a variety of different gastrointestinal functions including peristalsis and mucous secretion. These four major cell types, alongside other cells with immune function such as M-cells and the Tuft cells [39, 68], support rapid and efficient digestion and absorption of ingested nutrients.

The structure of the crypt provides an ideal niche for the intestinal stem cell. The multipotent CBCCs reside at the bottom of the crypt giving rise to differentiated progeny that migrate away from the crypt base toward the gut lumen, Figure 1.1. The activity of Wnt in the crypt base is critical to normal gut development and maintenance [33, 41, 55, 73] and is thought to be provided by the surrounding Paneth cells in the small intestine and the Crypt Base Secretory Cells in the colon. Further study of the intestinal stem cell niche has revealed the crucial importance of Notch and Epidermal Growth Factor (EGF) signalling in conjunction with Bone Morphongenic Protein (BMP) and ErbB signalling inhibition in the maintenance of stem cell properties and determination of cell fate decisions [85, 103, 106].

Though the work of Cheng and Leblond elegantly showed that CBCCs can display stem-like properties, accurately defining the stem cell population within the intestinal crypt has been a topic of intensive study.

In order to define a cellular population as possessing stem potential, the ability to self-renew and to produce all differentiated lineages of that tissue must be demonstrated. The gold standard for demonstrating this is through long-term *in vivo* lineage tracing from the population in question. The most experimentally robust intestinal stem cell marker to date is the Leucine-rich-repeat-containing G-protein coupled receptor (Lgr5). Lgr5 is a Wnt target gene strongly expressed within the crypt base in a spatially restricted manner [4]. Long-term lineage tracing from Lgr5+ cells produced distinctive ribbons of marked cells emerging from the crypt base. Furthermore, Lgr5 appeared to mark a rapidly cycling population of CBCCs that displayed increased resistance to radiation when compared with the quiescent cell population found at the +4 position. Until this discovery it had been hypothesised that a hallmark of stemness was quiescence [16] and the cells at the +4 position seemed the most likely candidate for the stem cell population of the intestine.

However, Lgr5 is not the only marker capable of producing clonal ribbons in the intestine. Hopx encodes an atypical homeodomain-containing protein and specifically labels cells at the +4 position along the whole length of the murine intestine. The Hopx+ population display quiescent properties but also have the ability to interconvert with the Lgr5+ population [95]. Bmi1, a polycomb-repressing complex 1 (PRC1) component, has been shown to be expressed in cells at the +4 position in duodenum and jejunum but not more distally in the ileum [84]. Their stem cell capacity can be demonstrated through lineage tracing despite Bmi1 and Lgr5 labelling two functionally distinct populations [108]. Incorporation of synthetic nucleosides, such as bromodeoxyuridine (BrdU), can be used an indicator of a proliferating cell. However, if a cell retains this mark in the long-term, named a BrdU-retaining cell, this is indicative of a quiescent cell state. Mouse telomerase reverse transcriptase (*mTert*) expression is the key regulatory step in telomerase complex activity and has been shown to co-label with BrdU-retaining cells at the +4 position [11]. This observation suggests high telomerase activity in quiescent stem cells within the crypt and *mTert* expression could be a valid marker of a population of quiescent stem cells. The above is not an exhaustive list of potential stem cell markers in the intestine but highlight some of the key discoveries of recent years. Other notable markers not discussed here include: Mcm2 has been shown to label a population of intermittently cycling crypt base cells [78]

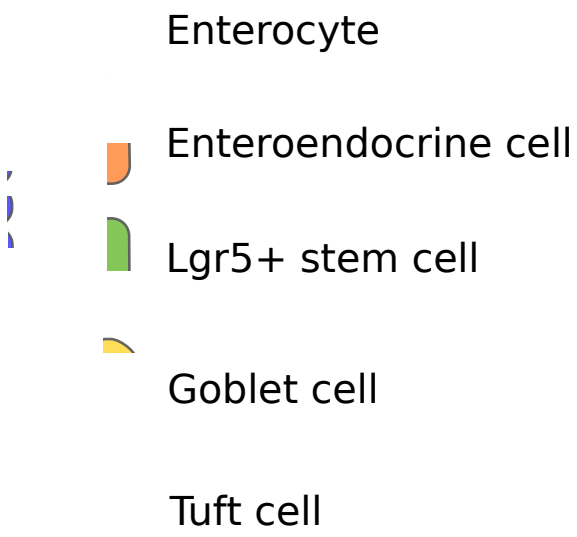

Fig. 1.1 Schematic of a single colonic crypt depicting the key differentiated cell types. The crypt base columnar cells are shown here as Lgr5+ cells.

and Lrig1 has been shown to be key in maintaining a stable intestinal stem cell population and labels a population of slowly cycling cells that can regenerate the intestinal epithelium following damage [77, 106]. Intestinal stem cell markers have been invaluable to the isolation, study and characterisation of stem cell state and the molecular mechanisms that govern stemness. However, the question still remains as to whether cell fate is a cell autonomous property or whether non-autonomous properties such as the location of the cell is more influential.

Recent studies suggest stemness in the intestinal crypt may be a cell non-autonomous property and that cellular plasticity is a pronounced feature of the crypt. One of the first cell markers in the crypt to display this plasticity was Delta-like 1 (Dll1). Dll1 is a notch ligand and is expressed by a subset of early stem cell daughters biased towards secretory lineage commitment. However, following irradiation, these cells are able to revert to a more primitive, multipotent state giving rise to all 4 major cell types of the intestine. Further work by Tetteh et al [97], used the Alkaline phosphate intestinal (Alpi) gene as a marker of well differentiated enterocyte precursors. During homeostasis Alpi+ cells are distinct from Lgr5+ rapidly cycling cells and Dll1+ secretory precursors and do not form clonal ribbons. However, when the Lgr5+ cell population is ablated, through engineered specific diptheria toxin sensitivity, Alpi+ cells are able to transdifferentiate to a stem-like state giving rise to both Paneth cell precursors and rapidly cycling CBCCs. This work provides strong evidence of stemness not being a 'hard-wired' property of cells within the intestinal crypt and instead factors such as contact with the niche having more influence on stem potential [30].

Due to the phenomenon of plasticity and interconversion of different cell types within the intestinal crypt, it appears that the existence of a single marker that specifically labels the stem cell population of the intestine may not be possible. Even a panel of markers may not adequately capture a pure intestinal stem cell population. Instead, a functional approach to defining the stem cell population of the intestine coupled with probabilistic calculations of stemness may yield more accurate definitions of what constitutes an intestinal stem cell.

An example of the potential of functionally defining stemness in the murine intestine is the work of Buczacki et al [12]. Through the use of a novel dimerisable dicreAB construct, lineage tracing only from cells that maintained histone protein H2B in the long-term, i.e. showed functional properties of a quiescent cell, is possible. It was observed that in the homeostatic state, quiescent label-retaining cells (LRCs) only contribute to the production of secretory progenitors and, therefore, do not display multipotency. However, during injury,

LRCs are able to revert to a multipotent state and contribute to epithelial maintenance. This work clearly shows that it is possible to define the role of stem cell populations based solely on the functional properties of those cells and without the need for a stem cell marker. Furthermore, due to the label retaining nature of these cells, it is also possible to isolate them through standard flow cytometry techniques for further downstream *in vitro* and molecular analysis. Novel techniques such as these have the potential to give clear insights into the functional and molecular characteristics of quiescent stem cells.

The dicreAB construct is a valuable tool in studying quiescent stem cells however, a different approach is required for studying the rapidly cycling stem cell population within the crypt base. Recent work by Kozar et al [54] propose a continuous labelling approach that utilises an unbiased genetic marker that can label all proliferating cells of the gut at random but at a low frequency. Only if a cell possesses true stem potential will it be able to retain this mark in the long-term (self-renewal) and give rise to differentiated daughter cells that contribute to epithelial maintenance (multipotency). Through quantification of crypt clone size, the authors were able to infer the functional stem cell number and stem cell replacement rate of homeostatic tissue and of dysplastic adenomatous tissue. Through this approach it was shown that the functional stem cell number is variable throughout the intestinal tract from 5 stem cells in the proximal small intestine through to 7 stem cells in the colon: far less than the 16 Lgr5+ cells seen in the crypt base [89]. This functional approach clearly shows that the gene expression stem cell markers currently used are only a surrogate of stem cell potential and do not only mark cells contributing to epithelial maintenance. This work is discussed further in Section 1.3.2.

1.2 Intestinal stem cell dynamics in health and disease

The continuing treadmill of differentiated cell types emerging from the crypt base before eventually undergoing apoptosis is driven by two key dynamics shown in Figure 1.2. Firstly, the rapid dynamics of the crypt-villus axis leads to differentiated cells moving away from the crypt base towards the gut lumen. Secondly, the slower dynamics of competition between equipotent stem cells in the crypt base leads to continual clonal conversion over time.

By understanding the rules by which intestinal stem cells replicate and maintain a stable population within a homeostatic setting, differences within the pathological setting can be inferred. One clear example of this is a dysplastic adenoma where glandular structures

morphologically and functionally resemble the homeostatic intestinal crypt, Figure 1.3. Just like the homeostatic crypt, the adenoma gland contains a clonal stem cell population at its base that continuously replicate giving rise to differentiated progeny. These progeny are then able to contribute to tumour repopulation and growth. By quantifying and comparing the stem cell dynamics of the homeostatic crypt with the adneoma gland, it may be possible to functionally and molecularly describe the processes that govern stem cell dynamics in the healthy and diseased states. Ultimately, through characterisation of these differences, it may be possible to design novel therapeutics that bias stem cell dynamics within the adenoma toward stasis or, ideally, regression.

1.2.1 The crypt-villus axis

The exact mechanism driving differentiated cell migration from the crypt base to eventually be shed into the gut lumen through apoptosis, either directly in the colon or via migration along the villi in the small intestine, is not known. A leading hypothesis is the 'co-operative model' proposed by Julian Heath [48] which proposes that the combined mitotic pressure provided by continuous cell division in the crypt base in combination with cell cohesiveness, cytoskeletal structure and underlying mesenchymal cell activity force the differentiating cells up the crypt-villus axis. Regardless, the rapid nature of this process is remarkable. The average transition time, in mouse, of a differentiated cell from crypt base to being shed into the gut lumen is 2-4 days including a rapid transit-amplifying state that lasts 2 days and leads to 4-5 cell divisions [102]. Paneth cell migration is of note as it is in the only cell type to escape the upward migration in the crypt and instead migrates toward the crypt base where it forms a niche for the intestinal stem cells.

1.2.2 Neutral stem cell competition and niche succession

Early observations of clonal patterning in the intestine were achieved through ^{3}H-thymidine labelling of phagosomes within the CBCCs. Cheng and Leblond observed these labelled phagosomes being passed onto all 4 differentiated cell types from the CBCC population leading them to postulate the 'unitarian theory of the origin of the four epithelial cell types' [24]. It was only in the late 1980s that the first conclusive evidence for a single cell of origin within the adult intestinal crypt emerged. Studies conducted in mice chimaeric for H-2 antigen and the *Dolichos biflorus* agglutinin (DBA) marker, showed that crypts either

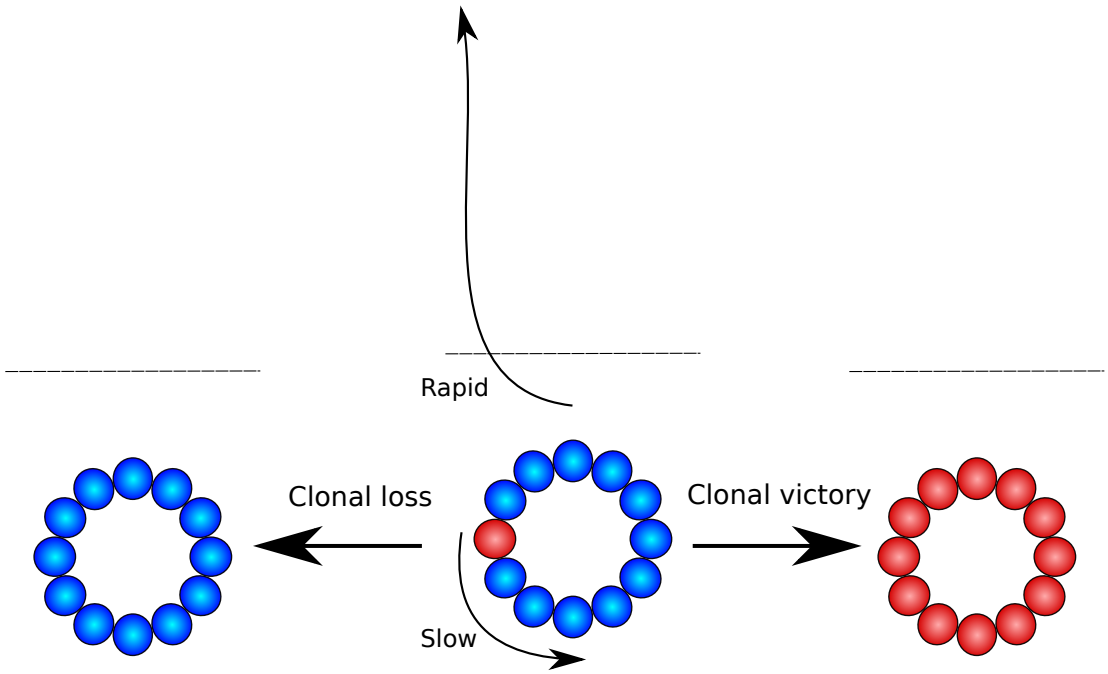

Fig. 1.2 Cellular turnover within the crypt is driven by two key dynamics. Top row represents a longitudinal section through the crypt base. Bottom row represents a transverse section through the crypt base at the level of the crypt base columnar cells. Centre panel represents a hypothetically marked single cell in the crypt base. The rapid dynamics of the crypt-villus axis generates a ribbon of differentiated cells above the marked cell within 2-4 days, in mouse. Simultaneously, the single marked cell is competing for space within the crypt base niche. This process is far slower and can lead to either the marked cell being lost from the niche leading to an unmarked crypt, as shown in the left panel, or clonal conversion of the crypt, shown on the right panel, leading to all cells in the crypt being marked.

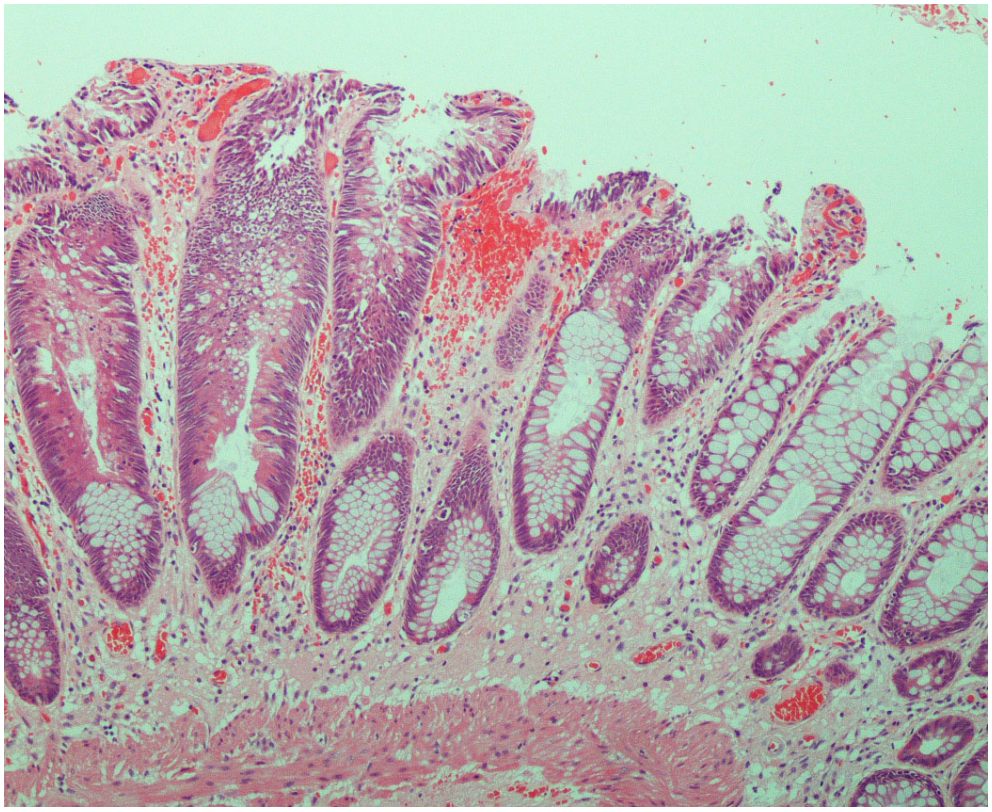

Fig. 1.3 H&E stained section of normal human colonic epithelium (right) showing homeostatic crypts and dysplastic adenoma (left) showing enlarged adenoma glands.

fully expressed H-2 antigen or the DBA marker thus inferring that each crypt must descend from a single progenitor [75].

Further studies observing loss of function at the *Dlb-1* locus, which determines expression of the binding site for the lectin DBA, demonstrated the utility of DNA mutation in observing niche succession. The approach required crossing of C57BL/6J mice that are *Dlb-1*$^{+/+}$ for DBA binding site expression in the intestinal epithelium with SWR mice that are *Dlb-1*$^{-/-}$ for DBA binding site. Therefore, in the F_1 cross, all offspring will be *Dlb-1*$^{+/-}$ needing only a single inactivating hit in the *Dlb-1*$^{+}$ locus to generate negative DBA peroxidase staining on a DBA positive background [104]. The natural mutation rate of *Dlb-1* locus is relatively low thus requiring the observation of many mice in order to attain meaningful results. Alternatively, administration of the mutagen ethylnitrosurea (ENU) leads to a burst of mutations causing a massive increase in the number of DBA negative clones, in essence generating a pulse-chase experiment. From this data, it was again possible to infer that all crypt progeny descend from a single progenitor cell. However, the observation of segmental negative staining clones suggested a process of niche succession was at work with the presence of an intermediate progenitor. Furthermore, little could be said about the exact location of this mutant progenitor based on this data [105]. Further work would be required in order to quantify the process of niche succession and a novel technique would be required to determine the exact location of the intestinal stem cell population.

The use of female mice hetereozygous for reduced expression of the X-linked histochemical enzyme, glucose-6-phosphate dehydrogenase (G6PD) has also been used as a clonal marker. The carcinogen dimethyl hydrazine (DMH) was used to induce somatic mutations. However, the authors did not observe any segmentally mutated crypts and, therefore, concluded that there must be one master stem cell of the crypt [42]. Similar analyses using X-inactivation alone, particularly in human tissue, has been confounded by the patch sizes generated by X-inactivation earlier in development [67]. Furthermore, no information about the precise location or nature of the intestinal stem cell could be inferred from this data and the question of the intestinal stem cell location persisted.

The average number of stem cells within the crypt needs to remain fixed throughout the life-time of an organism so as to prevent stem cell pool exhaustion or overgrowth. However, the observation of niche succession using the Dlb-1 reporter suggests that continuous cellular replacement is a central component of crypt homeostasis. In order to reconcile this apparent paradox a system of asymmetry must exist within the intestinal stem cell pop-

ulation. This could be achieved through two different processes. Firstly, a mechanism of intrinsic asymmetry whereby each daughter of a stem cell division is predetermined to either continue as a stem cell or to move down the lineage hierarchy and form a transit amplifying progenitor. Alternatively, a mechanism of population asymmetry could exist whereby lateral competition between neighbouring stem cells leads to one stem cell forming two stem cells whilst its neighbour differentiates to form two transit amplifying cells.

It was almost two decades after the initial observations of niche succession in the *Dlb-1* mutants that the process was fully described. Using low dose tamoxifen, an unbiased marking system was utilised to mark single, actively dividing cells in approximately 2% of all crypts throughout the gastrointestinal epithelium. By observing the crypt clone size change over time, it was possible to model the most likely mechanism of stem cell renewal within the intestinal crypt. This approach showed that the intestinal stem cell population replaced itself through a majority of symmetrical stem cell divisions consistent with more than one equipotent stem cell within the crypt. Furthermore, it was possible to show that this replacement rate was closely linked to the cell cycle time within the crypt base [59]. This work produced strong evidence in favour of an equipotent stem cell population central to intestinal homeostasis that maintained itself through primarily symmetric cell fate decisions.

At the same time, an alternative technique was developed to study the same mechanism. A *Confetti* cassette related to the Brainbow 2.1 construct was targeted to the *Rosa26* allele[58, 89]. When induced by Cre recombination, one of four different coloured reporter genes was expressed at random within the target cell. The authors restricted recombination to the Lgr5+ population through targeting of the Cre recombinase ERT fusion to the *Lgr5* locus coding region [4]. Through the use of short-term and long-term observations of clonal evolution using the multicolour lineage tracing strategy, it was again shown that the stem cell population within the crypt replaced themselves through symmetric fate decisions of neighbouring equipotent stem cells [89]. Together these two studies showed that the adult intestinal stem cell replacement followed a pattern of neutral drift [59, 89].

1.3 Lineage tracing

Lineage tracing experiments involve the labelling of single cells before observing their clonal evolution over time. Through co-staining of single clones with differentiated cell markers, it is possible to demonstrate the differentiation potential of that cell. Furthermore,

by observing long-term clone production from that single cell, it is possible to demonstrate self-renewal potential. Thus, the generation of a ribbon of differentiated cell types from the crypt base is the gold standard for demonstration of a cell with multipotent properties within the intestinal epithelium. Additionally, by observing the size of the clone over time it is possible to infer the dynamics that govern the replacement and maintenance of the stem cell population as discussed in Section 1.2.2. To date, the vast majority of lineage tracing experiments have been within model organisms due to the need for genetic modification and often the need for induction of reporters through tamoxifen, or similar compounds. The most commonly used methods for *in vivo* lineage tracing in the intestinal epithelium has been marker-based lineage tracing and more recently an alternative continuous labelling strategy has been described [54].

1.3.1 Marker-based lineage tracing

The most common lineage tracing experiment is based upon induction via a putative stem cell marker. Within the gut, the Leucine-rich repeat containing G-protein coupled receptor 5 (Lgr5) has been the most universally accepted intestinal stem cell marker to date. Through expression of a tamoxifen-inducible Cre recombinase under the *Lgr5* promoter it has been possible to induce fluorescent protein expression in Lgr5 expressing cells, and all descendants of that cell. As discussed in 1.2.2, the use of lineage tracing from this marker has allowed inference of the mechanism of cell replacement within the intestinal crypt. Furthermore, through flow cytometric isolation of $Lgr5^{hi}$ cells based on the level of fluorescent protein expression, it has been possible to increase the efficiency of clonal expansion of single intestinal stem cells into intestinal organoids [85, 86]. These experimental observations, emphasise the utility and value of accurate stem cell markers. A schematic depicting the generic marker-based lineage tracing approach is shown in Figure 1.4.

However, these experiments potentially lead to the labelling of overlapping populations with varying stem cell potential. It also relies upon the assumption that all Lgr5+ cells have equal stem cell potential and contribute equally to cell repopulation within the gut epithelium. As stem cell potential is likely to be a continuous property and certainly not a static nor binary state, tracing in such a manner is intrinsically biased and may not represent the true stem cell population. In order to achieve more accurate quantification of stem cell dynamics, an unbiased, functional approach to observing clonal evolution is required.

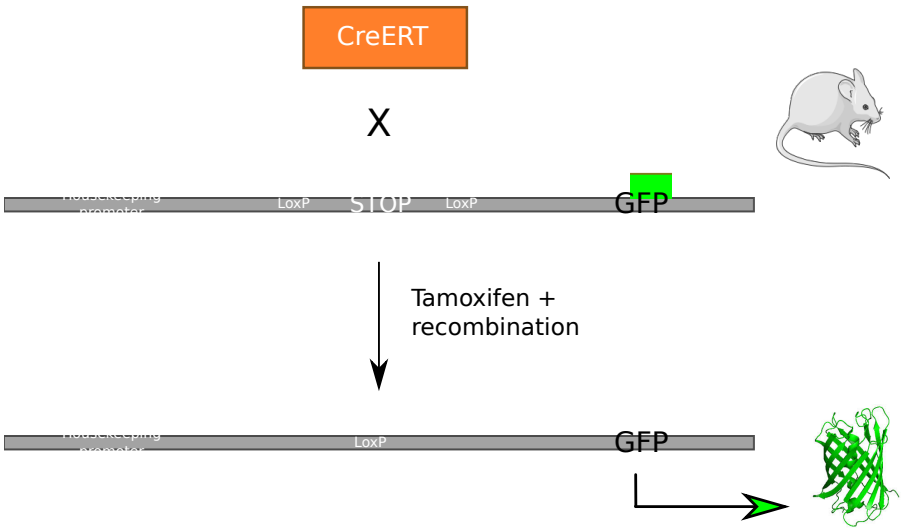

Fig. 1.4 Diagrammatic representation of the commonly used marker-based lineage tracing strategy. An inducible Cre recombinase (CreERT) is targeted to a putative stem cell marker promoter. Mice with this construct are then crossed with mice containing a reporter gene targeted to a housekeeping promoter contiguous with a transcriptional stop cassette flanked by LoxP sites.

1.3.2 Continuous labelling lineage tracing

Recent efforts have been focussed towards developing marker free approaches to infer stem cell dynamics within the intestinal crypt. By removing the need for a putative stem cell marker and instead lineage trace based on functional aspects of stem cells, any bias in the observation is removed. Kozar et al utilised a novel transgenic system in mice to express a histologically detectable protein at random within any proliferating cell of the gut [54]. This was done by introducing a reporter gene under a housekeeping promoter but contiguous with a $[CA]_{30}$ microsatellite that shifted the reporter gene out of frame. Through continual cell replication, errors are introduced into the highly mutable $[CA]_{30}$ microsatellite that bring the reporter gene in frame, Figure 1.5. If a single cell incurs an in-frame mutation and gives rise to any progeny, they will also inherit the in-frame mutation. Only cells with multipotent capacity will develop a ribbon of differentiated progeny in the intestinal epithelium. The ability to maintain a clone in the long-term, which could become a clonal victor in ongoing niche succession, requires self-renewal capacity. Therefore, any long-term clones observed must have the definitive characteristics of a stem cell: multipotency and self-renewal capacity. This strategy presents an unbiased, functional assay for studying clonal evolution within the intestinal crypt.

By observing the clonal evolution of marked crypts in a continuous labelling experiment, it is possible to infer stem cell replacement rate and functional stem cell number. Two basic metrics are required to calculate this:

1. The frequency of partly populated crypts (PPC) in mice at different ages (Figure 1.6). At any time during the lifetime of the mouse, there will be a small population of crypts that have undergone partial conversion and are equally as likely to drift to whole crypt conversion as they are to drift towards clonal extinction. The number of PPC remains constant throughout the lifetime of the mouse and is directly proportional to the average number of functional stem cells per crypt.

2. Wholly populated crypts (WPC) accumulate linearly with age (Figure 1.6). The rate at which whole crypt conversion occurs is directly related to the average rate of stem cell replacement.

The rate of both WPC accumulation and PPC frequency are also related to the rate of mutation within the $[CA]_{30}$ microsatellite.

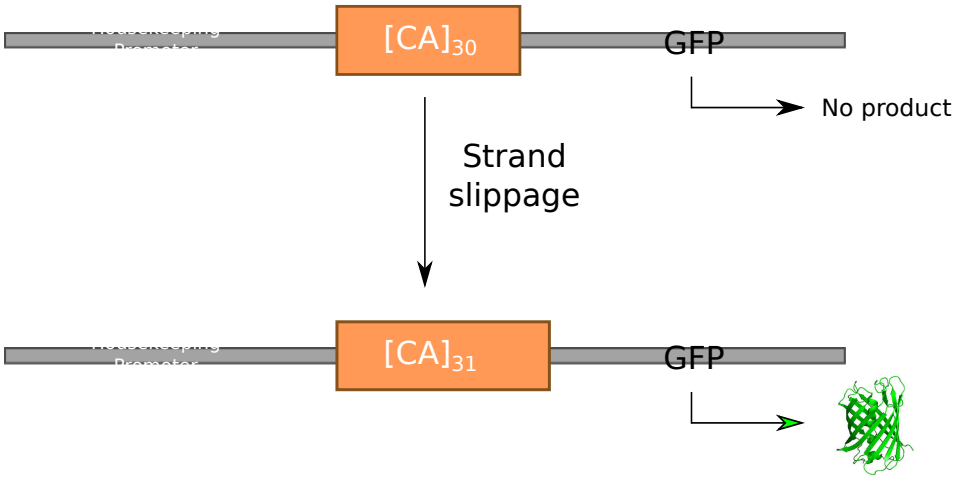

Fig. 1.5 Diagrammatic representation of the novel transgenic system used by Kozar et al [54] to randomly label cells within the murine gut. A reporter gene is targeted to a housekeeping promoter but is shifted out of frame by a $[CA]_{30}$ microsatellite. Through replication dependent strand slippage, the reporter gene can be brought in frame leading to expression of the reporter protein.

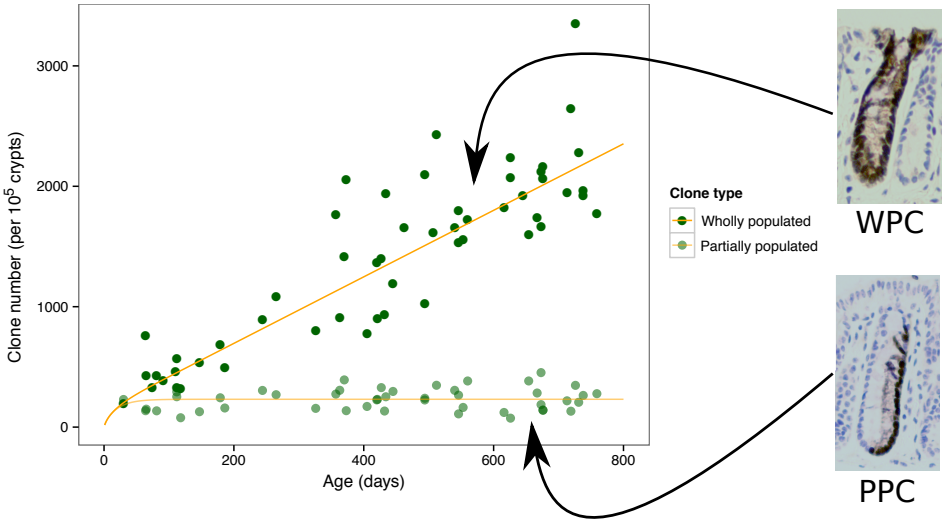

Fig. 1.6 Figure adapted from Kozar et al [54]. Plot depicts quantification of WPCs and PPCs in murine colon at different ages. Images on the right show immunohistochemically stained crypts from $[CA]_{30}$:eYFP mouse.

Using the model described above, it has been possible to quantify the stem cell dynamics within the murine intestinal crypt. The number of functional stem cells per crypt has been shown to be less than expected with around 6 stem cells per crypt. Furthermore, the functional stem cell number varies along the intestinal tract with an average of 5 functional stem cells per crypt in the proximal small intestine through to 7 per crypt in the colon. Additionally, it was shown that the rate at which an individual stem cell is replaced is also much slower than previously thought with an average of 0.1 per day in the proximal small intestine through to 0.3 per day in the colon. The data fitted well with a neutral drift pattern of stem cell replacement within the intestinal crypt thus regional difference in neutral drift dynamics can be explained by functional stem cell number and replacement rate differences alone. A similar analysis within adenoma glands also reveals smaller stem cell populations (approximately 9 stem cells per adenoma gland) and very high rates of stem cell replacement (each stem cell replaced approximately every day). This suggests that the majority of stem cell divisions within an adenoma are used to replace stem cell loss with very few clonal victors driving tumour growth and repopulation [54].

The findings of Kozar et al have far reaching implications for the study of intestinal stem cell biology in health and disease. Firstly, the number of *de facto* stem cells within the crypt is 50-70% lower than the number of Lgr5+ cells (16 Lgr5+ cells of which between 5 and 7 are functional stem cells). Though Lgr5 faithfully marks cells with stem cell characteristics in homeostasis, there is redundancy to the marker and not all Lgr5+ cells possess stem cell capability. Furthermore, the slower than expected stem cell replacement rate suggests that not all stem cell fate decisions are symmetric and there must be a level of asymmetric or asynchronous cell fate decisions within the stem cell population. Finally, the observation of a remarkably high stem cell replacement rate within the adenoma gland suggests a very high level of symmetric stem cell fate decisions in the dysplastic setting. Thus, in the dysplastic setting, the majority of stem cell fate decisions must be symmetric to maintain the stem cell population whilst the minority of cell fate decisions give rise to committed progenitors.

The observation of approximately 16 Lgr5+ stem cells per crypt contradicts the inference of 5-7 functional stem cells using unbiased continuous labelling. This paradox was somewhat reconciled with the development of intra-vital imaging techniques that allow *in vivo* imaging of various internal organs in the mouse [82]. Through the insertion of an abdominal imaging window in Lgr5:GFP mice, it was possible to perform near real time imaging of intestinal stem cell dynamics *in vivo*. Using this technique, it was shown that the

3D spatial positioning of the Lgr5+ cells had a quantifiable effect on stem cell dynamics. The crypt base can be split into two rings termed 'central cells' located from cell 0 to the +2 position and 'border cells' located in the +3 and +4 position. Through intra-vital imaging of the crypt, it was shown that central cells are three times more likely to fully colonise a crypt compared to the border cells and gives reason as to why Lgr5+ cells show differential stem potential [81]. These studies demonstrate how the combination of advanced imaging techniques of stem cell markers combined with functional, unbiased labelling techniques allows for accurate quantification of intestinal stem cell dynamics and the cellular processes driving it.

The use of transgenic mouse models has been invaluable to our understanding and quantification of intestinal stem cell biology. Translation of these techniques to human samples would allow for direct comparison and improved understanding of human intestinal stem cell dynamics. By understanding the homeostatic intestinal stem cell dynamics in humans, it may be possible to compare with various pathological states including inflammatory bowel disease and oncogenic predisposition. The techniques discussed thus far require transgenic manipulation of the test organism or induction using tamoxifen, ENU or similar compounds. Efforts to develop an effective tool for quantification of human stem cell dynamics are discussed in the following section.

1.4 Existing approaches to lineage tracing in the human intestinal crypt

The majority of attempts to perform lineage tracing in human intestine have come from observing the stochastic loss of clonal marks within the stem cell pool and detection by immunohistochemistry or similar techniques. By following the clonal behaviour over time, it is possible to approximate stem cell dynamics within the human intestine. These approaches have many benefits such as amenability to high throughput screening of thousands of single crypts and relatively low cost of implementation. However, the validity of these proteins as truly neutral clonal marks has not been assessed and the techniques do not lend themselves well to observations of multiple clonal marks in a single tissue and therefore restrict these studies largely to singleplex assays.

One such staining technique is the mild periodic acid-Schiff (mPAS). mPAS selectively stains non-*O*-acetylated sialomucins. Most individuals in the human population contain *O*-

acetylated sialomucins within their colonic epithelial cells therefore the cells do not stain for mPAS. However, some individuals are homozygous for an unknown polymorphic autosomal gene responsible for *O*-acetylation and, therefore, stain positive for mPAS. Of interest in colonic clonal biology are those individuals heterozygous for this polymorphism that have a high chance of generating sporadic mPAS+ clones. Histological analysis of the colons of heterozygous individuals display sporadic mutant crypts, that are increased following radiotherapy, and can be used to show that there is a common stem cell within the human colon. Furthermore, these observations have shown that the clone stabilisation time is greater than one year in humans compared to four weeks in mice. The difference in clearance time is probably due largely to a difference in stem cell number, crypt kinetics and the rate of crypt fission [17]. This technique is of great interest for the characterisation and quantification of intestinal stem cell dynamics in the human gut. The challenges lie in: 1) accumulating a large enough bank of tissue with heterozygote individuals and 2) identifying the polymorphic locus involved in mPAS+ staining and ensuring that loss of function does not effect normal stem cell dynamics.

Methylation patterns provide an alternative clonal mark that could be used to trace clones in human tissue. Restricted bisulphite sequencing of loci presumed to be neutral revealed increased methylation mosaicism with age and single crypt heterogeneity of methylation consistent with multiple adult stem cells replacing each other through regular symmetrical divisions [66, 109]. The clone stabilisation time was also shown to be far longer in humans than in mouse at over 8 years and perhaps as long as 15-40 years for some crypts.

An alternative approach is the use of mutations in the mitochondrially encoded cytochrome C oxidase (CCO) which leads to a biochemical defect that can be detected by the loss of enzyme-linked immunohistochemical staining [96]. This technique was applied to human tissue to infer the stem cell dynamics of the colonic crypt. Due to the low number of samples obtained for this study, the findings relied upon the assumption that the "wiggle" in the resulting ribbon of progeny from the crypt base represents a natural history of stem cell competition. Using this method, an estimate of 6 functional stem cells per crypt was inferred. This number is similar to the number of functional stem cells inferred in the murine crypt despite the 8-fold difference in crypt size and 2-fold difference in the mean number of cells in the crypt base circumference.

Furthermore, using inferences from the wiggle data, the authors suggest that stem cell competition in the crypt base follows a pattern of neutral drift described by the same one-

dimensional random walk model used to describe murine intestinal stem cell dynamics. The average crypt base to gut lumen transit time in humans is estimated to be 82-100 hours [57, 76] with the cell cycle time within the crypt base to be in the order of 24-48 hours [76]. Given that the observed wiggles, show multiple changes in clonal make-up, it seems unlikely that the wiggle solely constitutes a record of stem cell competition in the crypt base and most likely captures cell dynamics within other areas of the crypt such as the transit-amplifying compartment, where the majority of mitotic activity occurs within the crypt. The true utility of wiggle data in determining stem cell dynamics within the intestinal crypt remains to be seen.

Nonetheless, using observations of wholly converted CCO- crypts, an overall increase in CCO- patch size was seen in APC deficient tissue when compared with normal colonic epithelium. A further increase in fission rate was observed in dysplastic adenomas [2]. This ties in with previous observations suggesting crypt fission may be the key driver in mutation spread in the gastrointestinal epithelium [40]. This highlights the utility of CCO marking in determining clonal patterning of crypt units and epithelial sheet as a whole, both in the homeostatic and dysplastic setting.

There are also some caveats to using CCO as a clonal mark. Firstly, as this is a mitochondrially encoded gene, there are 1000s of copies of the gene in each cell and the threshold at which enough mutant copies are present to display a CCO- mark is unknown. It also relies upon the assumption that mutated mitochondial DNA (mtDNA) has an equal chance of propagation within the cell compared to wild-type mtDNA. Secondly, the effect of CCO deficiency on a cell is unknown. As the final component of the electron transport chain, it seems unlikely that the loss of CCO would not lead to perturbed cellular metabolism, as has ben reported in other systems previously [26, 79], the effect of which on stem cell competition is unknown. Regardless of the effect of CCO on the cell, a valid mouse model to assess the effect of CCO deficiency would be technically difficult due to the difficulty in stably editing the mitochondrial genome. Nonetheless, CCO has a valuable role in detecting clonally related single crypts and has already been applied to assessing the clonal relationship of distant and adjacent glands in adenomas [50].

Overall, the current repertoire of *in vivo* lineage tracing techniques in the human intestinal crypt remains limited and their caveats substantial. Developing a clonal mark that has the ability for high throughput screening and *a priori* knowledge of neutrality is essential to accurately quantifying the stem cell dynamics of human gastrointestinal tissues.

1.5 Utilising microsatellite sequencing for lineage tracing in the human intestinal crypt

As a result of studies into human crypt cell dynamics, it has emerged that intestinal stem cells undergo far more divisions than previously expected; approximately one division every 2-3 days [76]. This leaves the population susceptible to mutation due to replication error as well as large scale chromosomal rearrangements and aneuploidy. Furthermore, observations in mouse show random chromosome segregation during intestinal stem cell genome replication thus leaving the genome susceptible to accumulation of these mutations [29, 87]. The tissue architecture segregates individual clones thus somewhat reducing the impact of deleterious clones. However, mutation accumulation cumulatively increases the stem cell population's risk of oncogenic transformation over time.

Genomic changes within the crypt can predispose the tissue to malignant transformation but the majority of genomic changes are likely to have little, if any, effect on tissue homeostasis. These neutral genomic changes can be used as mark of the natural history of cellular lineage within the crypt and could be used to detect age-related clone size change. Kozar et al showed that microsatellites can be used as reporters of clonal lineage (albeit using a transgenic microsatellite contiguous with a reporter). This method could be translated to humans by isolating single human crypts, extracting the genomic DNA and sequencing endogenous microsatellites as a means of quantifying the size of mutant clones. Due to the relatively high mutation rate in microsatellites and no known phenotypic effects as a result of intergenic microsatellite length change, tracking endogenous microsatellite length change presents an ideal opportunity to track clone size in humans over time. By combining technologies associated with the amplification of low template copy DNA such as the polymerase chain reaction (PCR) with massively parallel DNA sequencing as described in Section 1.7, it should be possible to sequence microsatellites within a single crypt and detect mutations in endogenous loci. A highly simplified workflow is presented in Figure 1.7.

The vast diversity of microsatellite loci provides ample potential for multiplexed genotyping of cells to generate unique cellular 'barcodes' that could be used for clone size estimations and lineage heirarchy analyses. The potential for single cell genotyping based on microsatellite variability could be extended further when considering the biallelic presence of microsatellites and even higher levels of variability within longer repeats or in microsatellites more prone to mutation due to flanking region composition.

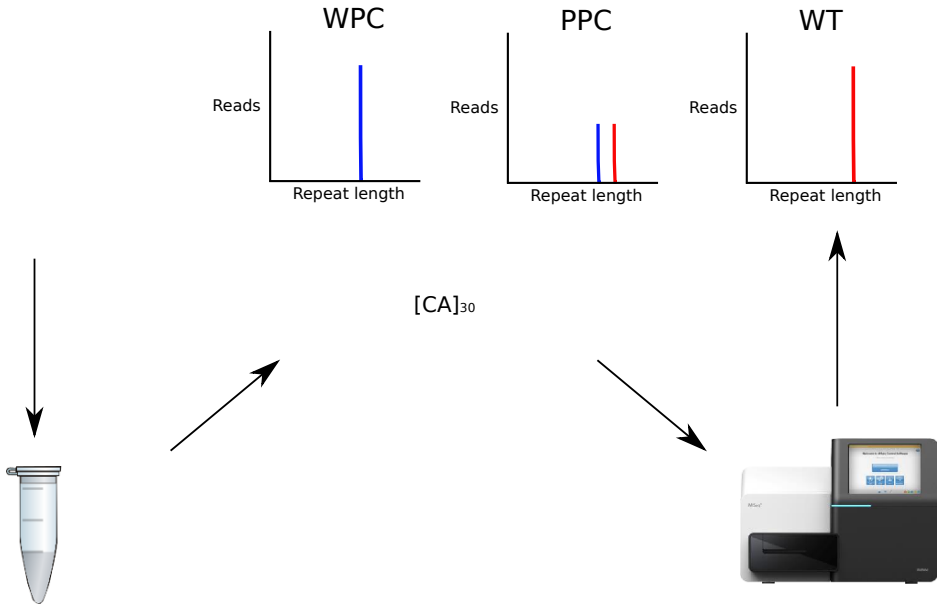

Fig. 1.7 Schematic showing the transgene-free approach to determining clone size in single intestinal crypts using endogenous $[CA]_{30}$ sequencing. Single crypts would be isolated using a micropipette, the DNA extracted and endogenous microsatellites amplified using primers that produce amplicons with sequencing adaptors at either end of the molecule. These amplification products would be prepared for Illumina sequencing and the data used to infer clonal status of the crypt. The data collected from this method would be amenable to the same analysis performed by Kozar et al [54] to infer functional stem cell number and stem cell replacement rate.

1.6 Microsatellites

Microsatellites are stretches of DNA that contain highly repetitive nucleotide units between 1 and 6 nucleotides in length ranging from 5 up to 50 contiguous units. Due to the highly repetitive nature of these regions they are highly prone to error due to polymerase slippage and deficient detection of mismatches by the mismatch repair pathway. As a result, regions containing microsatellites are more prone to mutation accumulation. Estimates of microsatellite mutation are in the region of 10^{-5} errors per cell division though this varies widely depending on tract length, nucleotide composition of the microsatellite and nucleotide composition of the flanking regions [3]. This rate is orders of magnitude higher than the germline mutation rate for single nucleotide variants estimated at approximately $2.3x10^{-8}$ per human generation, which takes into account multiple cell divisions between zygote and germ cell that is inherited by the next generation [65].

The rate of genomic mutation accumulation is variable between individuals, between tissues and across the genome. The spectrum of accumulated mutations leave a 'fingerprint' of single nucleotide variants (SNVs), insertions and deletions (indels) and large scale rearrangements, duplications and deletions. It has been estimated that 7% of non-coding and 14% of coding SNVs have phenotypic effects with larger scale changes hypothesised to have the potential for more significant effects [32]. Microsatellites have been demonstrated to be hypermutable due to polymerase error during replication, particularly in individuals with defects in mismatch repair pathways. It has been well demonstrated that large scale changes in these regions, particularly in coding regions, can lead to a range of pathological conditions such as Huntington's Disease, Fragile X Syndrome and Freidrich's Ataxia. However, small scale changes, outside coding regions are thought not to have any effect on gene expression or function.

Microsatellites, therefore, provide a unique opportunity to perform transgene-free, unbiased approaches to genotyping of restricted cell populations. Approximately 3% of the human genome at approximately 1 million unique loci, equivalent to 90M base pairs, is covered by microsatellites of varying composition and length [28, 93]. The majority of mutations within microsatellite regions are small scale and larger scale changes in microsatellite length are hypothesised to be as a result of smaller stepwise mutations. Large scale deletions of microsatellites are thought to be rarer events and lead to loss of surrounding genomic information as well as the microsatellite. Thus small scale changes in microsatellites could be used as a potential clonal mark at a high mutation rate with negligible effect on

cell function.

The rate of microsatellite change is inextricably linked to the length of the repeat. Through transgenic approaches, work in our lab determined that the $[CA]_{30}$ microsatellite provides a good balance between a high event rate coupled with a extremely low probability of reversion or doublet labelling within a single crypt. Through *in vitro* and *in vivo* estimation, the mutation rate at the $[CA]_{30}$ microsatellite has been shown to be approximately $1.1x10^{-4}$ mutants per mitosis [54] and estimations of similar mutations in $[CA]_8$ and $[CA]_{17}$ are $3.7x10^{-6}$ and $2.4x10^{-5}$ respectively [107]. These estimates clearly show the link between microsatellite length and mutation rate.

The mechanism of mutation for microsatellites is hypothesised to be as a result of insertion-deletion loops leading to skipping of repeat units; this process can lead to insertion or deletion of repeat units [46, 107]. This process is summarised in Figure 1.8. The insertion-deletion loops are proposed to occur in series as described by the stepwise mutation model [69]. In the homeostatic setting, most slippage events are repaired through the mismatch repair (MMR) pathway. However, if the mismatch repair pathway becomes defective, the mutation rate at microsatellites is massively increased. This is seen in inherited MMR component defects such as Lynch Syndrome or in the neoplastic setting in microsatellite instable tumours.

The MMR pathway is essential for the repair of microsatellite mutations. The first description of mammalian homologues of the MMR pathway was made in *Sacchamroyces cerevisiae* initially indicating the importance of MLH1, MSH2 and PMS1 [92]. Later studies showed involvement of other components of the MMR complex with major contributions from MLH1, MSH2 and MSH6 [9] confirming the earlier observations with a novel implication of MSH6. The mismatch repair pathway is the key repair pathway for correcting insertion-deletion loop mutations [71]. Loss of components of this repair complex have now been definitively shown to predispose to the most common form of inherited cancer, hereditary non-polyposis colorectal cancer (HNPCC), which shows an autosomal dominant pattern of inheritance with over an 80% lifetime risk of developing colorectal cancer [91]. Defects in MLH1, MSH2 and MSH6 are the most commonly defective loci inherited in Lynch Syndrome. All three form core components of the complex involved in the recognition of base mismatching, along with MSH3 and PMS2 (shown in Figure 1.9). An MSH2/MSH6 heterodimer complex is the most common form of the complex used to detect and repair single nucleotide mismatches and small extrahelical loops. MSH2/MSH3 heterodimers are more

commonly seen during the detection and repair of larger extrahelical loops. The role of MSH2 in both small-scale and large-scale repairs explains the increased severity of the phenotype seen in humans and in mouse models [51, 91]. Taken together, it can be seen that knockout of MSH2 in a mouse model is an opportunity to raise the microsatellite mutation rate.

Analysis of the mouse reference genome (mm9), performed by our lab, revealed 254 $[CA]_{30}$ microsatellites spread across all chromosomes; 244 on autosomes, 9 X-linked and one on the Y-chromsome. The availability of microsatellites around this length can be extended further by considering $[CA]_{29}$ and $[CA]_{31}$ microsatellites which extends the total number to 822 $[CA]_{29-31}$ microsatellites of which 793 are autosomal, 27 are X-linked and 2 are on the Y-chromosome. Analysis of the human reference genome, again performed by our lab, revealed far fewer $[CA]_{30}$ microsatellites; totalling only 10 across all chromosomes. When extending the analysis to look for $[CA]_n$ microsatellites between $[CA]_{28}$ and $[CA]_{32}$ there are a total of 63 microsatellites and between $[CA]_{25}$ and $[CA]_{35}$ there are a total of 395 microsatellites. A comparison of mouse and human $[CA]_n$ microsatellites is shown in Figure 1.10. These results are consistent with previously published results describing significantly fewer microsatellites in primate populations compared with rodent populations [28].

1.7 Next Generation Sequencing

Massively parallel DNA sequencing, more commonly termed Next Generation Sequencing, enables high throughput, high depth sequencing of DNA for a multitude of different analyses. Many forms of technology exist that utilise different methods for determining nucleotide sequence. As a result each technology possess differing benefits from longer reads through to improved detection of epigenetic modifications. Illumina's proprietary sequencing by synthesis (SBS) method [6], remains the leading technology used for most high throughput sequencing projects and, as such, will be the focus for this discussion.

Library preparation for Illumina sequencing begins with addition of specific adaptors to short DNA molecules targeted for sequencing (typically <1000bp in length). Addition of adaptors allows for annealing of target molecules to complementary oligonucleotides bound to a glass 'flow cell'. In a process known as bridge amplification, each target molecule is isothermally amplified to produce a clonal patch of molecules known as a 'cluster'. Each

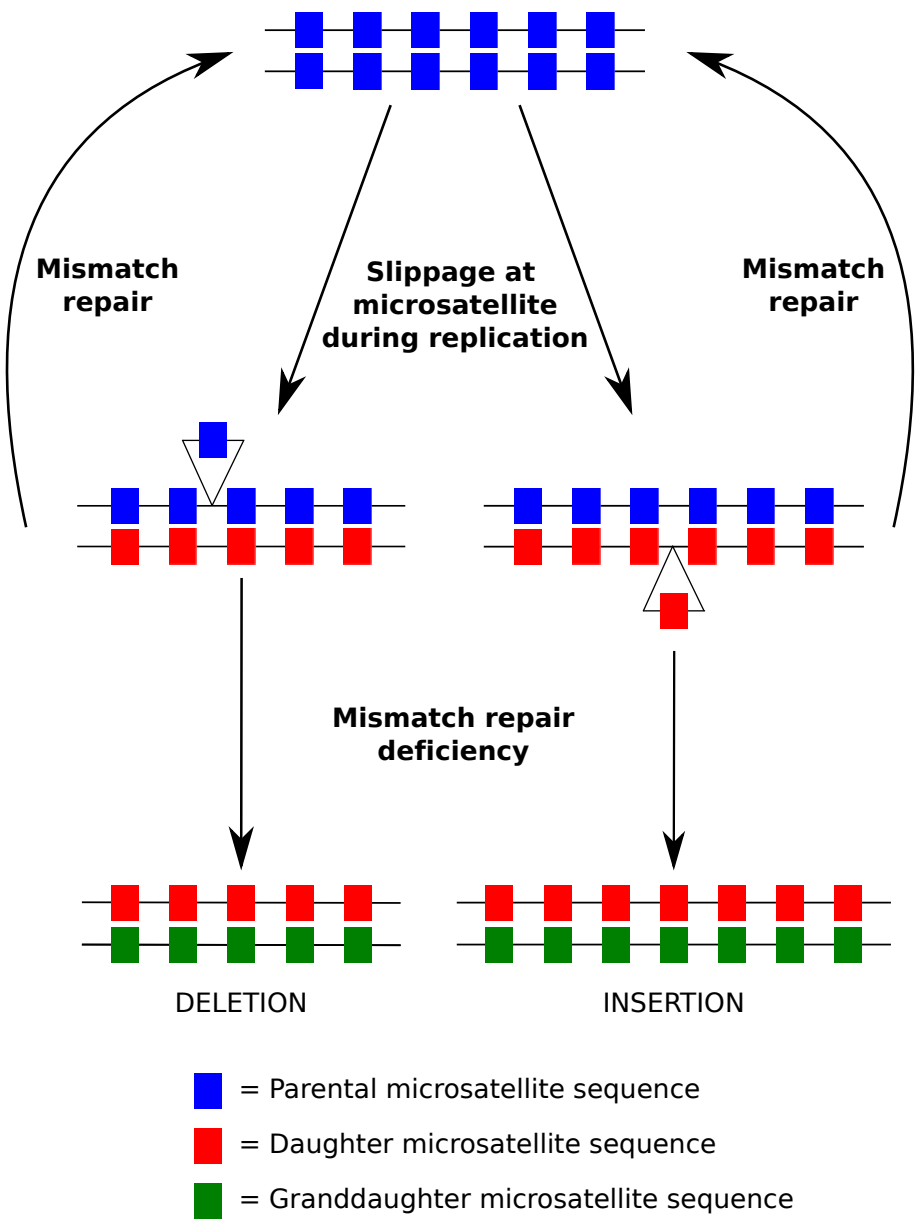

Fig. 1.8 Diagram depicting the hypothesised mechanism of loop insertion-deletion of microsatellite mutation. The mutation rate is massively increased by mismatch repair pathway deficiency though mutations can also occur as a result of stochastic infidelity of the mismatch repair pathway.

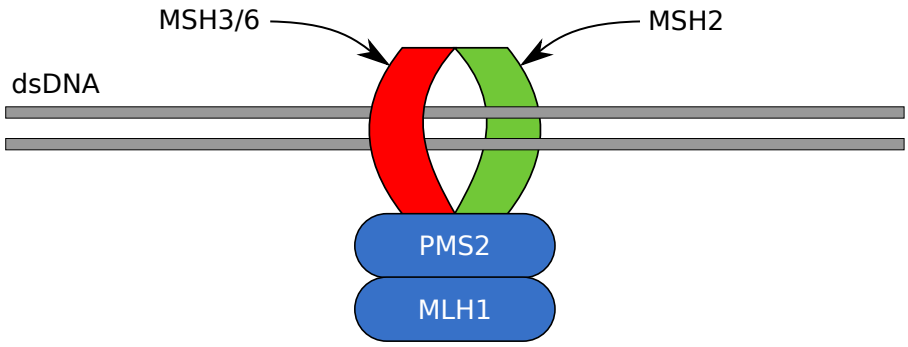

Fig. 1.9 Schematic depicting the core components of the mismatch detection and repair complex. MSH2, MSH3 and MSH6 form a ring structure around the DNA and can detect base mismatches and insertion-deletion loops. MLH1 and PMS2 form a ternary complex with mismatched DNA and possess the ATPase activity required for complex binding to DNA [51].

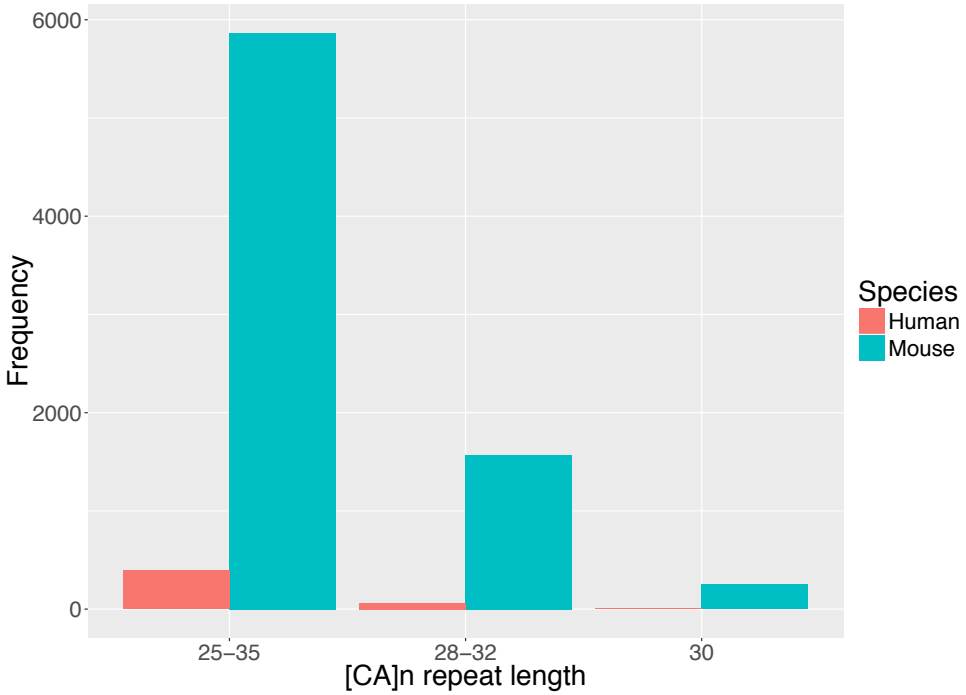

Fig. 1.10 Histogram displaying the number of $[CA]_n$ microsatellites at differing lengths in the mouse (mm9) and human (hg38) reference genomes including all autosomes, X and Y chromosomes. The abundance of microsatellites in mouse compared with the human genome is consistent with previously published results describing increased microsatellite frequency in rodent species compared with primates.

cluster is then denatured leaving only one strand of the target molecule attached to the flow cell. Through the use of a reversible chain terminating nucleotide, it is possible to add a single complementary nucleotide to the single stranded DNA before reversing the chain terminating group to allow the addition of another single nucleotide. Each time a nucleotide is added, it can be identified through a chemically conjugated, nucleotide-specific fluorophore. This process is repeated until up to 300 nucleotides are added thus giving read lengths of up to 300 base pairs for low read libraries (performed on the Illumina MiSeq) or up to 150 base pairs for high read libraries (patterned flow cells on the HiSeq 4000, NextSeq or HiSeq X Ten).

Furthermore, through the addition of adaptors at the 5' and 3' ends of the target molecule, it is possible to then repeat the sequencing process once more but from the opposite end of the molecule. This is known as 'paired-end' sequencing creating a forward and reverse read. This can be beneficial in two key ways: 1) for non-overlapping paired reads, it allows for more accurate read alignment to the reference genome and 2) for overlapping reads, you can generate two estimates of the nucleotides that are read twice thus increasing the confidence of each nucleotide call. The development of paired-end sequencing enables the generation of far more information from a sequencing library and can allow for more accurate read alignment and reference genome construction.

Due to the development of larger flow cells with more efficient sequencing chemistry, it has been possible to generate more reads for a given library far beyond what is necessary for sample analysis. This has motivated the development of indexing approaches that allow for the multiplexing of many sequencing libraries onto a single lane of the flow cell. Overall, this reduces the cost of sequencing, as the cost of sequencing a flow cell is split across multiple samples, and allows quicker turnaround, as more than one library can be sequenced each time a flow cell is run. To date, many indexing protocols have been developed allowing for the multiplexing of up to thousands of libraries onto a single flow cell. The use of such large indexing panels reduces the overall robustness of the pooled library as some indexes will perform better than others. However, very consistent and robust indexing panels containing 384 indexes have been developed and were used for this sequencing project.

1.8 Existing strategies for determining microsatellite length

Microsatellite sequencing is often termed short tandem repeat (STR) genotyping and is most commonly used in forensic sciences and population genetics. The most widely used technique for STR genotyping is multiplexed PCR followed by capillary electrophoresis (CE)-based separation. This technique is both time and cost-effective but lacks full sequence determination and is only semi-quantitative [10, 14, 15]. Therefore, use of this system for accurate quantification of mutant clone size from single crypts is not feasible.

Techniques utilising Illumina sequencing technology to ascertain microsatellite length have been developed with the objective of human identification [10, 111]. However, these techniques utilised PCR to generate 1000-1500 base pair amplicons before fragmenting and tagging the DNA in preparation for sequencing. Typically the simultaneous fragmentation and tagging process is done using a transposon-based system known as the 'transposome' [1, 94]. The use of a transposome requires a large amount of input DNA (at least 50ng); when compared with the DNA content of a single crypt (approximately 1.5ng per mouse crypt and 15ng per human crypt) is not a feasible means of library preparation. Furthermore, mixture experiments done using this technique, though often accurate, sporadically produced spurious results [10].

Other multiplexed, targeted sequencing protocols have been developed for genotyping purposes. A modified target enrichment protocol was used to pull down over 6000 STR loci before Illumina HiSeq sequencing however, a minimum of 2μg of high molecular weight human genomic DNA was required [44]. Molecular inversion probes have also been used to perform single molecule capture and molecular barcoding so as to account for PCR amplification bias during STR amplification. However, this protocol also requires large input amounts, this time 750ng of the *A. thaliana* genome, equivalent to approximately 2x10^6 genome copies [18]. The above gave ample reason to develop a custom quantitative STR genotyping protocol optimised for low template copy input.

In addition to the development of a protocol to generate accurate sequencing data from microsatellites, a computational method would also be required to analyse this data. Recently, computational pipelines such as lobSTR, RepeatSeq and STR-seq, have been developed that can utilise short-read, whole genome sequencing data to determine miscrosatellite length based on flank region mapping. Through the use of the STR-seq algorithm, a nine-fold increase in STR length calling error rate was observed in PCR-based library preparation methods when compared with PCR-free library preparation. Furthermore, analysis

of microsatellite length effect on mutation rate showed a direct correlation, as expected [37, 45, 49]. Similar algorithms have previously been developed for application to population genetics in Drosophila [34]. Aside from presently being financially prohibitive for this project, whole genome library preparation protocols currently require large input template requirements and often requiring whole genome amplification prior to preparation. However, once reliable whole genome library preparation is available for low template input, without the need for whole genome amplification, and the cost of sequencing reduced further, a computational pipeline could allow for genome wide genotyping of clones allowing the interrogation of multiple loci and achieving single cell resolution in clone size estimates. For the purposes of this project, custom tools for measurement of microsatellite length and inferring clone size from microsatellite sequencing data were developed, as described in Chapter 4.

As can be seen from the above, the development of a PCR-based Next Generation Sequencing solution for STR genotyping optimised for low template copy input has potential use beyond that of a neutral mark in areas such as forensic science and population genetics.

1.9 Project outline

In this dissertation, I describe the development of a novel, transgene-free method for intestinal crypt clone size estimation through multiplexed sequencing of endogenous microsatellites. In addition, I describe the computational method for determining clone size using microsatellite sequencing data as well as provide *in vitro* and *in vivo* validation of the protocol. This validated protocol was then used to observe expected age related clone size changes in murine colon. Finally, I present proof of principle data showing the application of this method to human tissues.

Chapter 2

Materials and methods

2.1 Animals

2.1.1 C57BL/6

Animals used for ageing studies were members of the C57BL/6 strain. Many of these animals were retired from existing breeding programmes so genotyping (Section 2.2.1) was performed to ensure that the relevant transgenes were not present. Animals were then aged in the Cancer Research UK Cambridge Institute Biological Resource Unit (BRU) until required.

2.1.2 Rosa26-$[CA]_{30}$-eYFP

The Rosa26-$[CA]_{30}$-eYFP mouse contain a dinucleotide repeat with a contiguous enhanced yellow fluorescent protein (eYFP) reporter gene targeted to the Rosa26 locus as previously described by Kozar et al [54] and depicted in Figure 1.5. Rosa26-$[CA]_{30}$-eYFP mice were subsequently maintained on the C57BL/6 background in a hemizygous state.

2.1.3 Villin-CreER2

The Villin-CreER2 construct contains a 9kb regulatory element of the *Villin* gene contiguous to a tamoxifen-inducible Cre recombinase construct. On induction with tamoxifen, Cre recombinase activity is induced throughout the intestinal epithelium as described by el Marjou et al [27]. The Villin-CreER2 mice were maintained as heterozygotes on a C57BL/6 background.

2.1.4 Msh2 flox

The Msh2 flox mouse contains LoxP sites targeted to either side of exon 6 in the *Msh2* locus. Recombination is driven by Cre recombinase leading to loss of exon 6 and Msh2 function. These mice were obtained from Dr. Paul Potter at MRC Harwell (MGI: 101816). The Msh2 flox mice were maintained in a homozygous state on a C57BL/6 background.

2.2 Genotyping and $[CA]_{30}$ Sanger sequencing

2.2.1 Transnetyx genotyping

Genotyping of the C57BL/6, Villin-CreER2 and Msh2 flox mice was outsourced to Transnetyx Inc. (Cordova, TN, USA). Ear biopsies for genotyping were loaded into a 96-well plate and shipped directly to Transnetyx where genotyping was performed using a Taqman-based assay. A full list of the primers used for genotyping can be found in Table 2.1.

Target	Forward primer	Reverse primer	Probe
Msh2 WT	GCGCTGGGTTGTTAATGGAAAT	CCTTCAGAGGGTGTGGAAACATTAATT	CAAAGCCACTAGAATAAG
Msh2 Flox	CCCGTGTCTCAAAATCTCTGATGT	CAGAGGGTGTGGAAACATTAATTATGACT	CCACTCTTGTGCAATGTA
eGFP	CGTCGTCCTTGAAGAAGATGGT	CACATGAAGCAGCACGACTT	CATGCCCGAAGGCTAC
Rosa26 WT	CTCTTACACTAAGCAATAA-AGAAATAAAAATTGAACTTCT	CTGCAGACTTAGCTTTCAGCTTTG	ACTGCTAGCTTTACTTAACTTT
ER	GAGCTGGTTCACATGATCAACTG	AGAAGGTGGACCTGATCATGGA	AAAGCCTGGCACCCTC

Table 2.1 Primers used by Transnetyx for genotyping. The ER probe was used to detect the presence of the *CreER2* locus. The eGFP probe was used to detect the presence of the *eYFP* locus

2.3 Animal care

All animals were cared for in the Cancer Research UK Cambridge Institute Biological Resource Unit (BRU) according to UK Home Office guidelines.

2.4 Treatment of animals

2.4.1 Tamoxifen administration

Tamoxifen was administered intraperitoneally to the Msh2$^{fl/fl}$;Villin-CreER2;Rosa26-[CA]$_{30}$-EYFP mice at 4mg/kg.

2.5 Intestine dissection

Mice were killed by cervical dislocation. Dissection was carried out through a ventral, midline incision. The small intestine was initially separated distal to the pyloric sphincter and dissected along the mesenteric border down to the ileocaecal junction. Caecum and colon are then dissected along the mesenteric border as far as the anus. Once dissected out, the first 3cm of proximal small intestine is discarded. 3cm to 15cm of proximal small intestine (12cm length) is then isolated. The distal 12cm of the small intestine closest to ileocaecal junction is then isolated. The caecum is separated from the colon. The middle small intestine section and caecum are discarded. The 12cm proximal intestine, 12cm distal intestine and colon are flushed with cold PBS using a blunt 10ml syringe.

2.6 Long-term storage of intestinal tissues

Occasionally, tissue was stored long-term at $-20\,^{\circ}$C by first processing as described in Section 2.5 before flushing the tissue with RNAlater stabilization solution (Thermo Fisher, AM7020). The flushed tissue was then transferred to a 50ml centrifuge tube and suspended in 25ml of RNAlater. The suspended tissue was then stored at $-20\,^{\circ}$C until required.

Once required, the tissue was removed from RNAlater and transferred to a petri dish containing PBS at room temperature. The tissue was flushed with room temperature PBS using a blunt 10ml syringe. Once the tissue had lost its rigidity, it was used for downstream processing and analysis.

If the tissue was to be used for fractionation, it must be noted that the crypt layer detaches as a single layer in fraction 1 and 2 following storage in RNAlater. Fractions containing crypt layers were centrifuged at 1500rpm for 5 minutes then the layers were broken up by gently pipetting with 20ml of cold PBS until a single crypt suspension was formed. The total volume was made to 50ml with cold PBS to dilute any cell free DNA present.

2.7 Crypt isolation

Crypt isolation is performed using a process known as 'fractionation'. Fractionation is the term used to describe the detachment of the epithelial cell layer from the underlying stroma. As differentiated cell types including villus structures detach in early stages of the protocol whilst crypts and lower layers of the intestinal epithelium detach later, the protocol allows for 'fractioning' of the different cell types for separate analysis. The protocol relies upon the incubation of the intestinal tissue in an alkaline EDTA solution before mechanical shearing. However, due to the structural differences in the specimens isolated in human and in mice, slightly differing techniques are required, as detailed below.

2.7.1 Murine intestinal tissue fractionation

Intestinal crypts were isolated from murine tissues based on a method developed by Bjerknes and Cheng [7]. 1.86g of Na_2EDTA and 500μl 10M NaOH were added to 500ml HBSS w/o Ca^{2+} or Mg^{2+} (Gibco). Proximal, distal and colon segments were everted so the luminal epithelium faced out and fed onto a 4mm diameter glass rod spiral. Each spiral was then submerged in 50ml of the pre-warmed modified HBSS solution and pulsed twice using a vibrating stirrer (Chemap AG, model CH-8604) at a high frequency with low amplitude so as to shake off the epithelium. The HBSS was then discarded. The spirals were re-submerged in 50ml pre-warmed HBSS solution and incubated for 10 minutes at 37 °C. Each spiral was then given 10 short pulses and the HBSS solution was then collected in a 50ml Falcon tube and immediately transferred to ice. This process was repeated five times giving a total of six fractions. Of which, the fourth and fifth fractions were enriched in single crypts.

2.7.2 Human intestinal tissue fractionation

Intestinal crypts were isolated from murine tissues based on a method developd by Fujimoto et al [36]. A 5-cm piece of normal colonic tissue was isolated from distal margins of patients undergoing resection for colon cancer at Addenbrooke's Hospital, Cambridge. The resections were performed by a consultant histopathologist to ensure important diagnostic specimens were not used and the tissue looked macroscopically normal. The specimens were incubated for 15 minutes in 0.04% sodium hypochlorite in PBS at room temperature to sterilise the surface. The colonic tissue was then washed in PBS and incubated in 3 mmol/l ethylenediaminetetraacetic acid (EDTA) plus 0.5 mmol/l dithiothreitol in PBS for 75 minutes at room temperature. The tissue was washed once with PBS, 10 ml of PBS was added, and the tube was shaken vigorously for 15 seconds. The crypts detach from the underlying mucosa during this shaking. The PBS containing the crypts was transferred to a 50ml centrifuge tube, and fresh PBS was added to the colonic tissue. The shaking step was repeated 4 times until the crypt yield diminished. The centrifuge tubes containing the crypt suspension were topped up to 50ml with cold PBS and kept on ice until ready for downstream processing.

2.7.3 PBS wash of crypt fractions

The single crypt enriched suspension was pelleted at 1200rpm for 2 minutes. Once pelleted, media was removed and the crypts were re-suspended in 50ml cold PBS.

2.7.4 Paraformaldehyde fixation of fractionated crypts

The same process as described in Section 2.7.3 is repeated twice. After the second pelleting, 4% PFA was then used to re-suspend the crypts and they were stored at 4 °C indefinitely.

2.8 Micropipette crypt picking

The following protocols can be performed using either a fluorescence or Brightfield dissecting microscope dependent on the requirements of the crypt isolation.

2.8.1 Alkaline lysis and neutralising buffer

The alkaline lysis buffer used to lyse single crypts was made as follows: 100mg NaOH was dissolved in 90ml of deionised water in an autoclaved 100ml bottle. 400μl of 50mM Na-EDTA was added before topping the solution up to 100ml.

The neutralising buffer used to neutralise the alkaline lysis buffer following single crypt lysis was made as follows: 484.4mg Tris-base was dissolved in 90ml deionised water in an autoclaved 100ml bottle. The solution was adjusted to pH 5 and topped up with deionised water to 100ml.

2.8.2 Standard technique

Under a low power dissecting microscope, 400μl of enriched crypt suspensions were spread on a siliconised (Sigmacote, Sigma-Aldrich, SL2-25ML) glass plate. Single crypts were then drawn into a micropipette using a mouth syphon. Drawing in small volumes of PBS interspersed with air optimised control of suction. Once drawn into the pipette, single crypts were then expelled into 3μl of alkaline lysis buffer. This was repeated a further seven times giving eight 3μl volumes of lysis buffer each containing one crypt. Each 3μl spot was then flooded with a further 3μl of alkaline lysis buffer and carefully transferred to an 8-well PCR strip.

2.8.3 Low adherence technique

As in Section 2.8.2 except single crypts were expelled directly into lobind PCR strips containing 6μl of alkaline lysis buffer. Observation under the microscope was used to confirm the transfer of the single crypt into the PCR strip for downstream processing.

2.8.4 Human crypt picking technique

The preparation was generated as in Section 2.8.2 except a hand held EZ-grip micropipette was used to isolate single crypts with a 170μm EZ-tip (both supplied by Research Instruments, UK). Crypts were transferred directly to lobind PCR strips as described in Section 2.8.3.

2.8.5 Crypt washing

Later experiments utilised crypt washing as a means of reducing the amount of cell free DNA contamination and diluting the amount of EDTA transferred into the final PCR reaction. For each crypt picked, 4 droplets of approximately 5μl of PBS were dispensed onto the siliconised glass plate. Once picked, each crypt was then moved from one droplet to the next whilst visually inspecting for any cell contamination. Finally, the crypt was transferred to the lysis buffer.

2.8.6 Crypt lysis protocol

Crypts in 6μl alkaline lysis buffer were then lysed via the following temperature cycling: 50 °C 3 hours, 75 °C 20 minutes, 80 °C 5 minutes, 4 °C hold. The alkaline lysis buffer was neutralised using an equal volume of neutralising buffer (6μl in this case). Crypt lysates were then stored at −20 °C for at least 48 hours before being transferred to −80 °C for long-term storage.

2.9 Polymerase chain reaction

All polymerase chain reactions were carried out on BioRad thermocycling blocks in either BioRad 8-well PCR strips (BioRad, TLS0801) or BioRad 96-well PCR plates (BioRad, HSS9601) sealed with BioRad Microseal 'B' adhesive seals (BioRad, MSB-1001). The same cycling conditions were used for each reaction condition: the key cycling conditions are outlined in Table 2.2. In this dissertation, the number of cycles is denoted in the PCR protocol name e.g. TS_66_35 is the TS_66 protocol with 35 cycles of PCR.

2.9.1 Standard Phusion reaction

Routine PCR was carried out using New England Biolabs Phusion High-Fidelity DNA polymerase (New England Biolabs, M0530S) using the high fidelity buffer. Buffer was diluted to 1x with dNTPs to a final concentration of 80nM and the required primer pairs to a final concentration of 400nM were added along with 0.2U of DNA polymerase. The required amount of template was diluted using nuclease free water to bring the whole reaction mix to 10μl.

2.9.2 Standard Q5 reaction

The Q5 high fidelity DNA polymerase (New England Biolabs, M0491S) was used to carry out some PCRs where indicated. The Q5 reaction buffer was diluted to 1X with dNTPs to a final concentration of 200nM and the required primer pairs to a final concentration of 400nM were added along with 0.2U of DNA polymerase. The required amount of template was diluted using nuclease free water to bring the whole reaction mix to 10μl.

2.9.3 Multiplexed Phusion reaction

Following multiple rounds of optimisation, the final mix for multiplexed Phusion PCR was as follows: NEB High Fidelity buffer diluted to 1x with dNTPs to a final concentration of 400nM, additional $MgCl_2$ to a final concentration of 1mM, pooled primer pairs to a final concentration of 0.25μM per primer pair, 1.6U of Phusion Hot Start Flex DNA polymerase (New England Biolabs, M0535S) made up to 10μl total volume with nuclease free water and template.

2.9.4 Indexing reaction

The Phusion High-Fidelity DNA polymerase (New England Biolabs, M0530S) was used to carry out indexing reactions. The high fidelity buffer was diluted to 1x with dNTPs to a final concentration of 80nM, either M13 of Fluidigm indexing primers to a final concentration of 0.1μM along with 0.2U of the Phusion DNA polymerase. Products of the initial TS_66 PCR protocol, were diluted 10x by adding 1μl of the reaction product to 9μl of nuclease free water and 4μl of the 10x dilution were added with nuclease free water to bring the whole reaction mix to 10μl.

2.10 PCR primers

All primers were synthesised by Sigma-Aldrich and purified by desalting. All vials were centrifuged at >10,000g for 30 seconds before dilution with nuclease free water to form a 100μM stock. Primers were stored at -20°C in a designated pre-PCR area. Before opening any primer stock, a UVP UV HEPA PCR system hood was UV irradiated at full power for 30 minutes and all handling was done within the hood.

2.10.1 Primer design

A custom Perl script was developed to identify the location of all $[CA]_{30}$ microsatellites within the reference mouse and human genomes. The output of this script also created a plain text file that was formatted for input into the online primer design tool BatchPrimer3 [110] with input of the 70bp sequence flanking either side of the $[CA]_{30}$.

The default BatchPrimer3 variables were used for all primer design except for stipulating that the forward and reverse primers must be no closer than 20bp and no further than 70bp from their respective end of the $[CA]_{30}$ tract. As a result, the maximum amplicon length was set at 200bp and the minimum amplicon length was set at 100bp with the optimal amplicon length stipulated as being 180bp. This ensures that the amplicon produced is suitable for paired-end 150bp sequencing and reduces amplicon length diversity within the library thus overcoming any significant length amplification bias.

2.10.2 NGS adaptors

Initially, the CS1/CS2 NGS adaptor system was added to the end of all primer pairs to allow compatibility with the Fluidigm barcoding system (Fluidigm, PN 100-3771) allowing pooling of up to 384 sequencing libraries. All primer pairs containing these adaptors will have their name prefixed with 'ADAP_' and will be indicated where relevant in the text. The sequence of these adaptors can be found in Table A.1.

In later experiments, an in-house designed NGS adaptor system was used named 'M13' adaptors. It was found that these adaptors reduced primer-dimer formation in multiplex PCR. These adaptors are compatible with an in-house designed barcoding system allowing the pooling of up to 384 sequencing libraries. The sequence of the indexing primers and the adaptor sequence can be found in Tables A.1 and A.2. All primer pairs containing these adaptors will have their name prefixed with 'M13' and will be indicated where relevant in the text.

2.10.3 Multiplex group design

All primers designed using the methods described in Section 2.10.1 were screened in duplicate and run on a 2% agarose gel to observe the presence of a single product at the expected size. All successful primers were taken forward for trial in multiplex PCR groups. The online tool MultiPLX 2.1 [52] was used to determine optimal multiplex PCR groups.

Each time a multiplex group was used in a sequencing experiment the balance of total reads for each amplicon was assessed. If an amplicon was found to have significantly higher total reads, the concentration of primer added to the initial pool was reduced. Equally, if an amplicon was found to have significantly reduced total reads, the concentration of primer added to the pool was increased. The final pooled concentration are indicated, along with the sequence of each primer pair, in Table A.3 for the mouse multiplex group and in Table A.4 for the human multiplex group.

2.11 Synthetic loci methods

2.11.1 Cloning of synthetic $[CA]_n$ loci

Endogenous loci were amplified from mouse genomic DNA isolated from tail samples using the Q5 polymerase (New England Biolabs, M0491S) as described in Section 2.9.2. The samples were cycled through the following conditions: 98 °C for 30 seconds then 60 cycles of 98 °C for 10 seconds, 59 °C for 20 seconds then 72 °C for 20 seconds before a final extension of 72 °C for 2 minutes. The primers used in this reaction are detailed in Table 2.3.

The product of the PCR reaction was cleaned and concentrated using the Zymo Research Clean and Concetrator-5 (Section 2.13.2) before being digested with BamHI (New England Biolabs, R0126S) and HindIII (New England Biolabs, R0104S) at 37 °C for 1 hour to create sticky ends for cloning. Digestion of the pUC19 plasmid (Addgene, GenBank: M77789) with BamHI and HindIII was also performed. The PCR product was ligated into the pUC19 plasmid using the Quick Ligation kit (New England Biolabs, M2200S).

Ligated plasmid was transformed into XL-1 Blue Competent Cells (Agilent, 200249) by heat shock. Transformed bacteria were grown on agar plates under ampicillin selection at 37 °C overnight. Colonies of surviving cells were picked and the plasmid purified using the Qiagen Spin Miniprep kit (Qiagen, 27106). Purified plasmid was screened by Sanger sequencing the pUC19 cloning region using M13 primers (Table 2.3) and the number of [CA] repeats counted manually. Sanger sequencing of this region was outsourced to Source BioScience (Nottingham, UK).

Selected clones at desired lengths were *in vivo* amplified further and purified using the Qiagen Plasmid Maxi kit (Qiagen, 12162). Final stocks of the plasmid were Sanger sequenced once more using Sanger validation primers (Table 2.3) and the number of [CA] repeats manually counted to ensure the length was as expected. Finally, the plasmids were

linearised as described in Section 2.11.2 and quantified using the Qubit dsDNA broad range assay (Section 2.14.1).

Type	Abbreviation	Step	Action	Cycles
Initial PCR	TS_66	1	95 °C for 10 minutes	
		2	95 °C for 15 seconds	Back to step 2 for 20-50 cycles
		3	66 °C for 30 seconds	
		4	72 °C for 1 minute	
		5	72 °C for 3 minutes	
		6	4 °C hold	
Barcode	BC_66	1	98 °C for 2 minutes	
		2	98 °C for 10 seconds	Back to step 2 for 10-15 cycles
		3	66 °C for 10 seconds	
		4	72 °C for 20 seconds	
		5	4 °C hold	

Table 2.2 Key PCR conditions used for PCR amplification from single crypt lysate.

Loci	Use	Forward primer	Reverse primer
a4_1365	Cloning amplification	GCGCTAGGATCCTCCCATGACTACTTCCTCCA	GGAGAGAAGCTTTCGCAAACTCAAATCCGGTG
s9_8328	Cloning amplification	GCGCTAGGATCCAACCCCTGTCTACCTTCAGC	GGAGAGAAGCTTCTCCCTCCCATGCCTCTATG
pUC19	M13 sequencing	TGTAAAACGACGGCCAGT	CAGGAAACAGCTATGACC
a4_1365	Sanger validation amplification	AACTCTATAAAACAACCCCTTCTGG	TATGTGCTCCCAGTGTGGAA
s9_8328	Sanger validation amplification	ACTTTAAGAGACAGGAGAAGCTC	GCTGGAGTCTGAGAACCACT
a4_1365	Sanger validation sequencing	AACTCTATAAAACAACCCCTTCTGG	TATGTGCTCCCAGTGTGGAA
s9_8328	Sanger validation sequencing	ACCTCAGCAAAGAAACTATGTCT	TGGCTGCAGTATTTCGCAAA

Table 2.3 Sequences of primers used for screening and validation of synthetic loci length.

2.11.2 Plasmid linearisation

For more accurate quantification of the pUC19 plasmid and better simulation of the endogenous loci *in vitro* state, the pUC19 plasmid was linearised using ScaI restriction enzyme (New England Biolabs, R3122S). The reaction mixture contained 20U of rectriction enzyme, 5μg of plasmid template and 1X NEBuffer made up to 50μl with nuclease free water. The samples were incubated at 37°C for 1 hour.

To avoid re-circularisation of the plasmid, the samples were treated with Calf Intestinal Alkaline Phosphatase (CIP) (New England Biolabs, M0290S). 2μl of CIP was added to each reaction sample and incubated at 37°C for 1 hour. Plasmids were visualised using gel electrophoresis (Section 2.12) to ensure linearisation had occurred and were quantified using the Qubit dsDNA broad range assay (Section 2.14.1).

2.12 Gel electrophoresis

Agarose gels were made from 1x TAE buffer and laboratory grade agarose with 4μl of Ethidium Bromide added per 100ml of agar. 10μl of amplified DNA was removed from test reactions and 2μl of 6x loading buffer added. Samples were then run on 2% agarose gels for 45-75 minutes at 90V alongside a DNA ladder (New England Biolabs, N0467S). Gels were visualised and photographed using a UVP ultraviolet transilluminated system.

2.13 DNA purification and concentration

2.13.1 DNA extraction

DNA was extracted and purified from tail samples or crypt pellets using the Qiagen DNeasy Blood and Tissue Kit (Qiagen, 69504) as per manufacturer's instructions.

2.13.2 DNA concentration

DNA clean-up and concentration post-PCR was performed using the Zymo Clean and Concentrator-5 kit (Zymo Research, D4004) as per manufacturer's instructions. For larger initial volumes of DNA suspensions, the Millipore Amicon Ultra-2 Centrifugal Filter Unit with Ultracel-30 membrane (Millipore, UFC203024) as per manufacturer's instructions.

2.14 DNA quantification

2.14.1 Qubit dsDNA broad range assay

Qubit dsDNA broad range assay (Thermo Fisher Scientific, Q32853) was used to quantify bulk DNA solutions at 10ng/μl or above as per manufacturer's instructions. Infrequently, Qubit dsDNA high sensitivity assay (Thermo Fisher Scientific, Q32854) was used to quantify low concentration bulk DNA solutions as per manufacturer's instructions. In both cases, the Qubit 2.0 fluorometer (Thermo Fisher Scientific, Waltham, MA, USA) was used with Qubit 500μl assay tubes (Thermo Fisher Scientific, Q32856) to ascertain readings.

2.14.2 Quant-IT dsDNA high sensitivity assay

Purified DNA samples were quantified on the PHERAstar DNA quantification system (BMG Labtech, Aylesbury, UK) using the Quant-IT dsDNA high sensitivity assay (Thermo Fisher Scientific, Q33120) as per manufacturer's instructions. This assay was used to ensure samples within a library were adequately balanced and allowed for approximate sample normalisation.

2.15 MagJET NGS library size selection

The MagJET system is designed to size select NGS libraries using a magnetic bead suspension in a 1.5ml microcentrifuge tube with the ability to size select a maximum of 8 samples at a time. In order to increase the throughput of the kit, all the volumes of solution suggested in the manufacturer's instructions were halved so that size selection can be done in a 1.2ml MIDI 96-well plate (Thermo Scientific, AB-0564) using the Ambion Magnetic Stand-96 (Thermo Scientific, AM10027). Using the manufacturer's instructions as a guide, the kit was further improved for maximal recovery in a 96-well format. The optimised protocol is described below.

2.15.1 Size selection calibration

Before size selecting a DNA amplicon pool, it is necessary to calibrate the size selection buffer volume to maximise product capture and minimise adaptor binding to the capture beads.

To begin, binding mix was prepared at the required volume in multiples of 1600μl using 576μl of 100% isopropanol and 1024μl of binding buffer. If the amount of binding mix exceeded 16,000μl, calibrated scales were used to measure out 4.53g of 100% isopropanol and 11.02g binding buffer per 16,000μl of binding mix. Two 1.2ml MIDI 96-well plates were prepared: the first with variable volumes of binding mix (recommended ranges in Table 2.4) and the second plate containing 50μl of binding mix per well. The magnetic beads were vortexed well and 2.5μl of magnetic bead suspension stock was added to each well of plate 1 and plate 2.

For size selection calibration, 50μl of fragment mix provided in the kit was added to each well of plate 1. The mixture was then shaken at 1500rpm for 15 seconds using the Illumina High Speed Microplate Shaker (BioSurplus, 3002322) before incubating at room temperature for 5 minutes. The plate was then placed on the magnetic plate and incubated at room temperature for a further 5 minutes. This step removed larger fragments by binding them to the magnetic beads: the target DNA is in solution.

The supernatant from plate 1 was transferred to plate 2 whilst keeping plate 1 on the magnetic plate. Plate 2 was vortexed on the plate shaker for 15 seconds at 1500rpm before being incubated at room temperature for 5 minutes. Plate 2 was then incubated at room temperature on the magnetic plate for 5 minutes. The desired product was bound to the magnetic bead and fragments smaller than the desired size were in solution.

Supernatant from plate 2 was discarded; plate was taken off the magnetic plate and add 15μl of elution buffer was added. Plate 2 was vortexed on the plate shaker for 15 seconds at 1500rpm and incubated at room temperature for 1 minute. Plate 2 was placed on the magnetic plate and incubated for 3 minutes. The supernatant was transferred to a fresh 96 well plate.

The products of the size selection were analysed on a 2% agarose gel or on the the Agilent Bioanalyser DNA high sensitivity kit (Section 2.16.1). The selected fragments were compared with the size of the library to be selected and an optimal binding mix volume was selected. This volume was used in the size selection protocol (Section 2.15.2).

2.15.2 Size selection protocol

To begin the size selection protocol, binding mix was prepared at the required volume in multiples of 1600μl using 576μl of 100% isopropanol and 1024μl of binding buffer. If the amount of binding mix exceeded 16,000μl, calibrated scales were used to measure out 4.53g

of 100% isopropanol and 11.02g binding buffer per 16,000μl of binding mix. In addition, 600μl of wash buffer were made per sample adding 3 parts 100% ethanol to 1 part wash buffer provided in the kit.

Three 1.2ml MIDI 96-well plates were prepared: the first with 350μl of binding mix; the second plate with the volume of binding mix determined in the calibration step to select the desired fragment range and the third plate containing 50μl binding mix per well. The magnetic beads were vortexed well and 2.5μl of magnetic bead stock was added to each well of plate 1, 2 and 3.

Exactly 50μl of sample was added to plate 1; samples with less than 50μl volume were topped up with nuclease free water. Plate 1 was then vortexed on the plate shaker for 15 seconds at 1500rpm; incubated at room temperature for 5 minutes before a 5 minute incubation on the magnetic plate. The target DNA was bound to the beads and shorter fragments were in the solution.

Plate 1 was kept on the magnetic plate and the supernatant discarded. Plate 1 was taken off the magnetic plate and 52.5μl elution buffer was added. Plate was then vortexed on the plate shaker for 15 seconds at 1500rpm then incubated at room temperature before a 1 minute incubation of the magnetic plate. 50μl of supernatant was transferred from plate 1 to plate 2. Plate 2 was vortexed on the plate shaker for 15 seconds at 1500rpm; incubated at room temperature for 5 minutes before being incubated on the magnetic plate for 5 minutes. Larger fragments were bound to the beads and the desired DNA fragments were in solution.

With plate 2 on the magnetic plate, the supernatant was transferred from plate 2 to plate 3. Plate 3 was vortexed on the plate shaker for 15 seconds at 1500rpm; incubated at room temperature for 5 minutes before being incubated on the magnetic plate for 5 minutes. The target DNA was bound to the beads.

Plate 3 was kept on the magnetic plate and the supernatant discarded. With plate 3 still on the magnetic plate, 300μl of wash buffer was added and incubated at room temperature for 30 seconds. The was buffer was discarded and the wash step was repeated for a total of two times with plate 3 on the magnetic plate at all times. The bead pellet was allowed to air dry for up to 5 minutes.

10-50μl of elution buffer was added to the bead pellet off the magnetic plate. Plate 3 was vortexed for 15 seconds at 1500rpm; incubated at room temperature for 1 minute before being incubated on the magnetic plate for 3 minutes. The size selected library was in solution. Samples were transferred to a fresh 96-well plate for storage or downstream

analysis.

2.16 NGS library quality control

Before submission, all NGS libraries were analysed to ensure the library was present at the expected size range with no adaptor-dimer contamination. The library was submitted to the Cancer Research UK Cambridge Institute Genomics core facility either between 10nM and 20nM or at exactly 4nM, as quantified by qPCR.

2.16.1 Agilent Bioanalyser library analysis

For products expected at low concentration (5-500pg/μl), the Agilent DNA High Sensitivity kit (Agilent Technologies, 5067-4626) was used. For products expected at higher concentrations (0.1-50ng/μl), the Agilent DNA 1000 kit (Agilent Technologies, 5067-1505) was used. In both instances, the kits were used with the Agilent 2100 Bioanalyser system (Agilent Technologies, Santa Clara, CA, USA).

2.16.2 qPCR quantification of NGS library

A serial dilution of libraries was first done to obtain 1 in 10,000 and 1 in 100,000 dilution of libraries quantified on the Agilent Bioanalyser (Section 2.16.1) to be between 10nM and 20nM. Each of these dilutions was done in triplicate.

The Kapa Biosystems qPCR library quantification kit for ABI Prism (Kapa Biosystems, KK4844) was used to quantify the libraries. 6μl of master mix (with added primer pair mix) was added to each required well of a 384-well plate including 24 wells for triplicates of 8 standards. 4μl of each replicate of the 1 in 10,000 and 1 in 100,000 dilutions were added to the master mix and 4μl of standard was added to the master mix in triplicate. The plate was sealed with an optical seal and spun down at 280g for 1 minute.

qPCR analysis was performed on the ABI 7900HT instrument (Fisher Scientific - UK Ltd, Loughborough, UK). qPCR cycling conditions were as follows: 95°C for 5 minutes then 35 cycles of 95°C for 30 seconds then 60°C for 45 seconds. An optional melt curve from 60°C to 95°C would be performed at this stage.

Threshold value for Ct was manually adjusted to fall within the exponential part of each Ct-fluorescence curve. The standard curve was checked to have a slope between -3.1 and

-3.4 and have an R^2 value of over 0.98 for log(fluorescence)-Ct. Any outliers to the standard curve were removed.

Exported data was analysed in a Microsoft Office Excel spreadsheet generated by the Cancer Research UK Cambridge Institute Genomics Core Facility. The spreadsheet used the exported data and standard curve to calculate the median Ct for each triplicate dilution of each library. A standard deviation was also calculated to flag any triplicates with a large variance around the median value. Any outliers within a triplicate were removed from downstream analysis. Using the median Ct value for both dilutions of each library and the standard curve Ct values along with average fragment size calculated from the bioanalyser, it was possible to infer the concentration of the initial input library.

Using the inferred concentrations, each library was diluted to 4nM before submission to the Cancer Research UK Cambridge Institute (CRUK-CI) Genomics Core Facility for Illumina MiSeq or HiSeq 4000 sequencing.

2.17 Computational methods

2.17.1 FASTQ file demultiplexing

FASTQ files generated by the Illumina MiSeq or HiSeq 4000 systems in the CRUK-CI Genomics Core Facility were demultiplexed by the CRUK-CI Bioinformatics Core Facility using the demuxFQ tool.

2.17.2 FASTQ quality filtering

The fastq_quality_filter (part of the FASTX toolkit: http://hannonlab.cshl.edu/fastx_toolkit/) was used to filter all reads within each FASTQ file only keeping reads that had 80% or more of the read at Q20 or above.

2.18 Digital PCR

The Fluidigm 12.765 digital PCR array (Fluidigm, BMK-M-12.765) was used to perform digital PCR as per manufacturer's instructions using the Fluidigm IFC Controller MX and Fluidigm Biomark HD instruments (Fluidigm, San Francisco, CA, USA). A probe for the

FoxD3 locus on chromosome 4 with a FAM reporter (TaqMan, Mm02384867_s1) was duplexed with a copy number reference assay probe for the Tfrc locus on chromosome 16 with a VIC reporter (ABI, 4458366). ROX acted as a passive control. Together these probes acted as quantitative indicators for the number of copies of each of those loci as they detect intra-genomic regions. The two loci should show high levels of correlation as these cells are highly likely to be in a diploid state. Therefore, the number of copies of each of these loci will be equivalent to the number of genome copies within each lysate.

Per reaction, 5μl of TaqMan Gene Expression Master Mix (ABI, PN 4369016), 0.5μl of 20X gene expression loading buffer, 0.5μl of FoxD3-FAM assay, 0.5μl Tfrc-VIC assay and 3.5μl of template DNA were added. The DNA template was first heated to 95°C for 1 minute before addition to the reaction to generate single-stranded DNA template and, therefore, increase the number of detectable molecules.

2.19 SYBR green qPCR assay

The Life Technologies SYBR Green PCR Master Mix (Life Technologies, 4309155) was used to identify any amplification bias between microsatellites of differing lengths. The kit was used as per manufacturer's instructions including an optimisation reaction to identify the optimal forward and reverse primer concentrations for each loci. The primers used for qPCR analysis had the CS1/CS2 sequencing adaptors attached and were first diluted to 5μM before addition to the reaction. Full details of the primers used can be found in Table A.3. For loci s9_8328, 0.12μl of forward primer and 0.72μl of reverse primer was added to each 12μl reaction. For loci a4_1365, 0.12μl of forward primer and 2.16μl of reverse primer was added to each 12μl reaction. These volumes were determined to be optimal through melt curve analysis and agarose gel electrophoresis visualisation of products generated from varying forward and reverse primer volume combinations.

Each reaction was performed in a 384-well plate with 6μl of SYBR master mix, 1.2μl of linearised plasmid at 1pg/μl, primers at the optimised volume and made up to 12μl total volume with nuclease free water. The analysis was performed on the Life Technologies QuantStudio 4 (Thermo Fisher Scientific Inc, Waltham, MA, USA) with the following cycling conditions: 95°C for 10 minutes then 40 cycles of 95°C for 15 seconds followed by 60°C for 1 minute. A melt curve was performed each time to check for any primer-dimer contamination.

Average desired DNA fragment length	**Binding mix, μl**
200bp	400μl , 450μl , 500μl , 550μl , 600μl
300bp	300μl , 350μl , 400μl , 450μl , 500μl
400bp	250μl , 300μl , 350μl , 400μl , 450μl
500bp	200μl , 250μl , 300μl , 350μl , 400μl
700bp	150μl , 200μl , 250μl , 300μl , 350μl

Table 2.4 This table was adapted from the MagJET size selection manufacturer's instruction manual. For a given desired fragment length, the range of binding mix volumes to be tested during size calibration is shown.

Chapter 3

Development of multiplexed microsatellite sequencing protocol

To quantify clone size distributions using microsatellite sequencing, a PCR protocol needed to be optimised to balance the requirements of: 1) faithfully reflecting the original distribution of microsatellite lengths, 2) amplify the original crypt DNA enough to meet sequencing concentration requirements and 3) ideally include multiple loci in a single reaction. Existing quantitative approaches to Illumina sequencing of microsatellites have not been optimised for low template copy input nor have attempts been to made to quantify and ameliorate any effect of *in vitro* polymerase slippage during microsatellite amplification using the polymerse chain reaction (PCR) prior to sequencing.

Based on the observations made by Kozar et al [54], calculations can be done to approximate the incidence of clones. In the colon of a 300 day old mouse, around 60 crypts would require sequencing in order to observe one WPC and 218 crypts in order to observe one PPC, Table 3.1. As the current method for crypt isolation takes a full day for one person to isolate 100-200 crypts, simply analysing one microsatellite will significantly limit the number of clonal events observed. By amplifying up to 20 microsatellites from a single crypt, it should be possible to observe more events and increase the power of the method. Therefore, a method utilising a multiplexed amplification and sequencing protocol to measure multiple microsatellites per crypt is required. However, the use of multiplexed PCR produces additional challenges to attaining adequate amplification from low template copies largely due to non-specific generation of primer-dimers and allele dropout.

A summary of the proposed pipeline for the library preparation of single crypts for mul-

tiplexed microsatellite sequencing is summarised in Figure 3.1. A single crypt suspension was obtained using existing methods of intestinal epithelium fractionation adapted from the method described by Bjerknes and Cheng [7] and fully described in Section 2.7. A mouth syphon attached to a micropipette was used to isolate single crypts under a dissecting microscope. The single crypts were then lysed in individual wells. The resulting lysate entered a multiplexed PCR to amplify endogenous loci before sequencing initially on the Illumina MiSeq before larger scale sequencing on the Illumina HiSeq 4000.

The proposed workflow appears simplistic but major technical challenges are associated with each step. The key challenges are discussed below.

3.1 Microsatellite sequencing from low template copies

Each intestinal crypt is a discrete clonal unit. The cell number per crypt is significantly higher in human colonic crypts at approximately 2000 cells compared to approximately 250 cells per mouse ilieal crypt [53] i.e. 4000 and 500 genome copies respectively. Amplifying hypermutable regions, such as microsatellites, from such low template copies renders the protocol susceptible to many sources of error and noise. The key sources of error are: 1) DNA contamination in crypt lysate, 2) *in vitro* polymerase slippage, 3) allele dropout and 4) sequencing error. Each of these sources of error are discussed below.

3.1.1 DNA contamination during crypt isolation

Initially, the isolation of single crypts must be validated to ensure the entire DNA content of the crypt is isolated without the transfer of contaminating DNA. In this context, contaminant DNA would include any DNA included in the analysis that has originated from anywhere other than short-lived daughter cells generated from the stem cell population at the base of the crypt. This would include stromal cells, immune infiltrate, long-lived Paneth cells, cells from adjacent crypts and cell free DNA transferred along with the crypt during the experimental protocol. Minimisation of contaminant DNA is imperative to ensure consistency and accuracy between samples. It is important that the degree to which contaminating DNA effects the accuracy and resolution of the protocol is determined. If this is shown to be an issue, steps will be taken to try and reduce this effect.

Crypt Status	Frequency	1x $[CA]_{30}$	8x $[CA]_{30}$	20x $[CA]_{30}$
WPC	1680 per 10^5 crypts	1 per 60 crypts	1 per 8 crypts	1 per 3 crypts
PPC	458 per 10^5 crypts	1 per 218 crypts	1 per 27 crypts	1 per 12 crypts

Table 3.1 Estimates of the number of WPCs and PPCs observed in multiplex groups of differing sizes. The estimates are adapted from Kozar et al [54] for colon of a 300 day old mouse. The estimates account for autosomal loci being biallelic but only include the rate for in-frame mutations thus likely represents a conservative estimate.

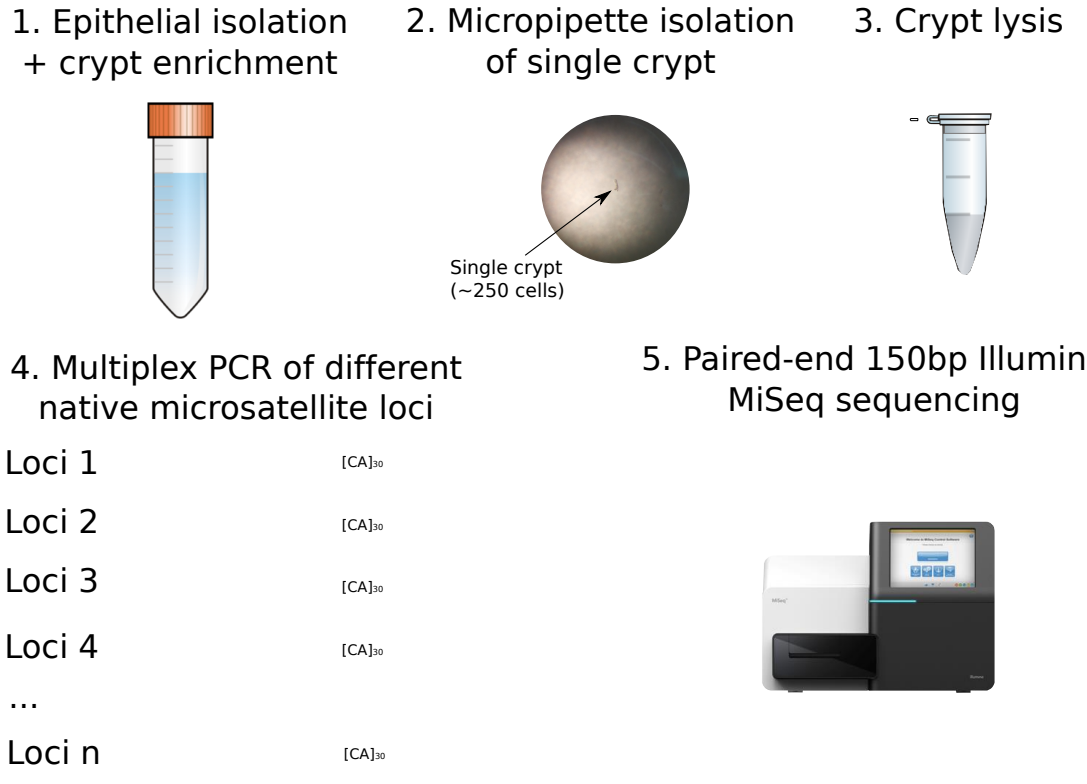

Fig. 3.1 A stepwise schematic showing the approach to sequencing multiple endogenous $[CA]_{30}$ microsatellites from a single intestinal crypt.

3.1.2 *In vitro* polymerase slippage

Key to the use of microsatellite sequencing as a means of quantifying clone size is generation of consistent microsatellite length calls with maximised signal-to-noise ratios. By amplifying a region containing a microsatellite, using PCR, the polymerase has no external mechanism for mismatch repair and relies solely upon the intrinsic replication fidelity and proofreading activity of the enzyme to detect and repair error whilst amplifying the microsatellite. As a result, the *in vitro* microsatellite mutation rate is thought to be far higher than the *in vivo* mutation rate. It has also been observed that the 'stutter bands' formed from PCR produce a bias towards contraction of microsatellites, particularly in dinucleotide and mononucleotide tracts [47, 64, 88]. As a result the sequencing data produced from the PCR amplified product is a 'blurred' representation of the parental copies of DNA. To gain sequencing data that accurately represents the actual microsatellite length and maximises the signal-to-noise ratio, the PCR protocol must be optimised to reduce polymerase error. Central to this is: 1) reducing the number of PCR cycles required to produce a sequencing library from single crypts; 2) screening many different DNA polymerases to determine polymerases with the lowest microsatellite mutation rate *in vitro*.

In addition to a PCR protocol for single microsatellite amplification, any method requires optimisation for multiplexing. Key to optimising a successful multiplex PCR assay is reducing the amount of off-target annealing events that occur leading to primer-dimer formation. Further, caution must be taken when using multiplex PCR assays in NGS protocols as primer-dimers are able to undergo barcoding reactions and generate sequencing reads on the Illumina platform. Due to their shorter length, they are able to form sequencing clusters far more effectively than the target amplicons leading to large losses of usable data from each sequencing run. Reduction of primer-dimer with the potential need for a size selection step therefore is crucial to developing this multiplexed NGS method.

The balance of representation of different locus specific amplicons must also be optimised. Particularly well performing primer pairs may amplify their target very effectively and lead to a massive over representation of that amplicon in the sequencing data. Thus, through an iterative process of sequencing and re-calibration, the concentration of each primer pair within the multiplex PCR reaction must be altered to balance the representation of all amplicons in each crypt amplification reaction.

3.1.3 Allele dropout

A commonly reported issue in any low template copy sequencing project is the phenomenon known as 'allele dropout'. This is a process by which stochastic non-amplification of certain alleles during early rounds of PCR leads to under-representation of that variant and over-representation of other variants [13]. If microsatellite sequencing is to be an effective method for quantifying clone size, it is necessary that the proportional representation of each allele is maintained throughout amplification. Previous reports have described methods for accounting for PCR bias and allele dropout using molecular barcoding: a process by which individual primer-template annealing events are tagged with a unique nucleotide sequence. During analysis, it is possible to account for over and under represented unique identifiers thus removing PCR bias. These techniques, though incredibly powerful, have only recently been optimised for multiplex use in low copy number samples [72, 90] so were not used in this project.

The polymerase selected for this project will also require high amplification efficiency so as to reduce the incidence of allele dropout in early cycles of PCR. Allele dropout rates will also be closely associated with careful primer design and titring of primer pair concentration so as to maximise annealing events whilst limiting primer-dimer formation and the occurence of false positives. Experiments were also performed to demonstrate the effect of allele dropout on clone size estimates.

3.1.4 Illumina sequencing of repetitive DNA

The final major technical challenge in this protocol is sequencing highly repetitive DNA on the Illumina sequencing platform. As the Illumina sequencing protocol requires an isothermal amplification step, further errors in microsatellite length may be generated. Furthermore, sequencing by synthesis is susceptible to a process known as 'phasing', this occurs when an additional nucleotide is added during synthesis leading to the read 'jumping' one nucleotide ahead. The combination of these two error modes is hard to detect when sequencing highly repetitive stretches of DNA and would be particularly accentuated when sequencing very simple repeats such as mononucleotide and dinucleotide repeats. At the time this project began, only one paper had been published showing that it may be feasible to use short-read Illumina technology to sequence microsatellites [10] but the longest microsatellite studied was a 26 unit tetranucleotide repeat and no dinucleotide repeats were studied. Therefore, when we set out to develop this protocol, the feasibility of accurately sequenc-

ing dinucleotide repeats using an Illumina sequencing platform was unknown. Subsequent studies have shown the utility of Illumina sequencing platforms in STR genotyping based on whole genome sequencing data [111] but did not use the Illumina platform for ultra-high depth, targeted sequencing of microsatellites. To the best of my knowledge, this work represents the first described targeted resequencing protocol for dinculeotide microsatellites on the Illumina sequencing platform.

3.1.5 Use of crypt equivalents

The eventual intention for this protocol is to sequence endogenous microsatellites within human tissues, however all of the data in this chapter was obtained from work with murine crypts and mouse reference DNA.

For the majority of optimisation experiments, crypt equivalents were used to simulate a single crypt lysate. These were made by diluting mouse reference DNA, extracted from ear or tail clippings, with the lysis buffer used for extracting DNA from single crypts. The dilution created lysis buffer containing mouse DNA at the concentration that would be expected if a single crypt were lysed in 12μl of buffer, as per the crypt picking protocol (Section 2.8). The same calculation was performed when optimising the protocol for human crypts. The steps used to calculate these concentrations are shown in Table 3.2. The use of either crypt equivalents or real crypt lysates will be indicated where relevant throughout this dissertation.

This chapter discusses the key optimisation steps taken and describes the optimal library preparation method for the multiplexed amplification of multiple $[CA]_{30}$ microsatellites from a single murine crypt.

3.2 Optimising the isolation of single murine intestinal crypts

Creation of a suspension of single crypts from mice can be accomplished by incubation of whole gut in alkaline, EDTA solution before applying mechanical shearing to detach crypts from the underlying stroma. This method is fully described in Section 2.7. Once in suspension, it is possible to isolate single crypts using a mouth syphon attached to a micropipette. An initial method of single crypt isolation was performed as described in Section 2.8.2. Through rational re-development, many aspects of the picking technique were altered to reduce the chance of losing crypt material; this re-designed protocol is described in Section 2.8.3 and was named the 'low adherence' technique. The two key changes were:

1) using plastics coated with DNA repellent so as to prevent DNA sticking to the plastic and being lost during transfer and 2) expelling picked crypts directly into the lysis buffer whilst visually inspecting the transfer through a dissecting microscope. In addition to improving DNA isolation from single crypts, the effect of 4% PFA fixation on crypt fractions was assessed with the aim to store the crypts long-term and allow multiple days of crypt picking to be performed. This was performed as described in Section 2.7.4. Crypts were then taken out of PFA and picked using the low adherence technique.

To determine the most effective method of crypt DNA isolation, digital PCR was utilised to accurately quantify the amount of DNA isolated by: the standard technique, the low adherence technique and following PFA fixation. Figure 3.2 shows the variable amount of DNA isolated using standard crypt capture techniques and the lack of DNA detection after 4% PFA incubation (none detected). However, using the novel low adherence technique, the reproducibly captured the genome equivalent of approximately 200-350 cells (250 cells = 500 genome equivalents = 1.5ng DNA) from multiple crypts. Therefore, all crypts used in this project were picked using this improved picking technique.

3.3 Quantifying and reducing cell free DNA contamination

During the fractionation process, cells can undergo lysis and apoptosis leading to the release of cell-free DNA (cfDNA). Due to the majority of crypts within the single crypt suspension containing wild-type $[CA]_{30}$ microsatellites, capture of this DNA in single crypt lysate will lead to under-estimates of mutant clone size.

To determine a baseline value of cell free DNA within the fraction media, the amount of cfDNA was quantified. To do this, mouse colon underwent the epithelium isolation and crypt enrichment protocol described in Section 2.7. The single crypt suspension in fractions 4 and 5 were then pelleted, re-suspended in PBS and left on ice. At time 0 and at 1 hour, 2 hours and 3 hours, the crypt suspension was re-pelleted and 2ml of the PBS was removed.

To observe the effects of re-suspending the crypt fractions in different volumes of PBS, two mice were used and fractions 4 and 5 from these animals were suspended in either 25ml or 50ml PBS before quantification. At one hour intervals, the crypts were pelleted and a 2ml volume of the PBS suspension was sampled. The PBS containing cfDNA was concentrated using the large volume concentrator, Section 2.13.2, and accurately quantified

	Mouse	**Human**
One genome	$2.7x10^9$ base pairs	$3.3x10^9$ base pairs
One base pair	650Da	650Da
Mass one genome	$1.8x10^{12}$ Da	$2.2x10^{12}$ Da
Weight one genome	3.01pg	3.67pg
# genomes in one crypt	500	4000
DNA content of one crypt	1.5ng	14.7ng
Concentration in 12μl buffer	0.125ng/μl	1.2ng/μl

Table 3.2 Table displaying steps taken to calculate crypt equivalent concentration.

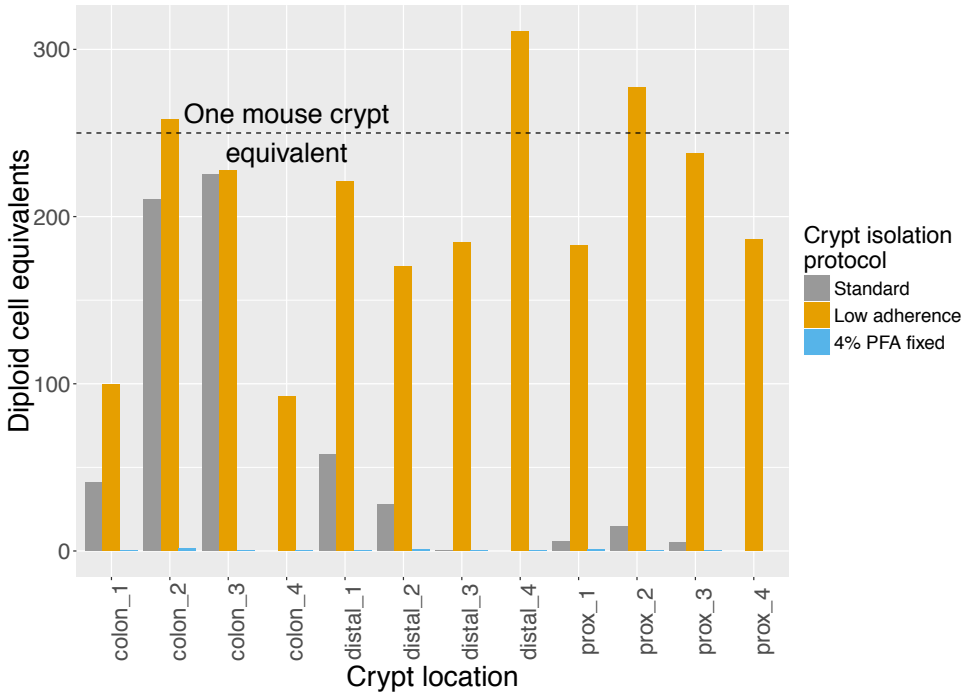

Fig. 3.2 Histogram summarising digital PCR quantification of DNA isolated by standard techniques, an improved low adherence technique and after 4% PFA fixation from murine crypts in the proximal small intestine, distal small intestine and colon. Based on this data, it was decided to only use the low adherence technique for all future crypt picking experiments.

using the Qubit dsDNA HS kit, Section 2.14.2. The results of this quantification are shown in Figure 3.3. As expected, re-suspending in 50ml of PBS rather than 25ml of PBS has a significant effect on reducing the concentration of cfDNA seen in the fractionation media. Furthermore, fraction 5 appears to contain less cfDNA when compared with fraction 4. This is likely due to the reduced crypt density seen in fraction 5. The amount of cell free DNA increases with time and is likely a result of cell lysis. Crypts should, therefore, be picked as soon after fractionation as possible. From these results, only fraction 5 was used for crypt picking after re-suspension of the crypts in 50ml PBS. This condition also displays the least change in cell free DNA concentration with time allowing a larger time frame for picking.

3.4 Minimising cell debris contamination

Single cells become dislodged from crypts and surrounding stroma during fractionation and are visible throughout the single crypt suspension. These single cells can be accidentally isolated along with the single crypt. Due to the majority of these single cells containing wild-type $[CA]_{30}$ microsatellites, isolation of these single cells along with the crypt will again lead to an under-estimate of the mutant clone size.

As there is no efficient way of removing single cell debris or clumps of cell debris from the single crypt suspension, the most effective way of minimising this effect is through visual inspection of the crypt as it is being picked. By passing the crypt through droplets of PBS before final transfer to the lysis buffer, it is easy to visualise and exclude single cells during transfer. Furthermore, this method allows the amount of cfDNA and EDTA content of the media to be minimised thus further improving downstream analysis.

3.5 Amplifying endogenous $[CA]_{30}$ loci

Primer pairs were designed as described in Section 2.10.1. A key aspect of the primer design is the addition of a 20-70 nucleotide flanking sequence at the 5' and 3' ends of the $[CA]_{30}$ tract. Sequence complexity within these flanking sequences are key to highly specific primer binding and amplicon generation. From a technical perspective, high sequence complexity in the first 5 nucleotides of the amplicon is particularly important for adequate calibration of the Illumina sequencing instrument. If sequence complexity is low early in the read, this can effect the sequencing quality for the remainder of the sequencing run.

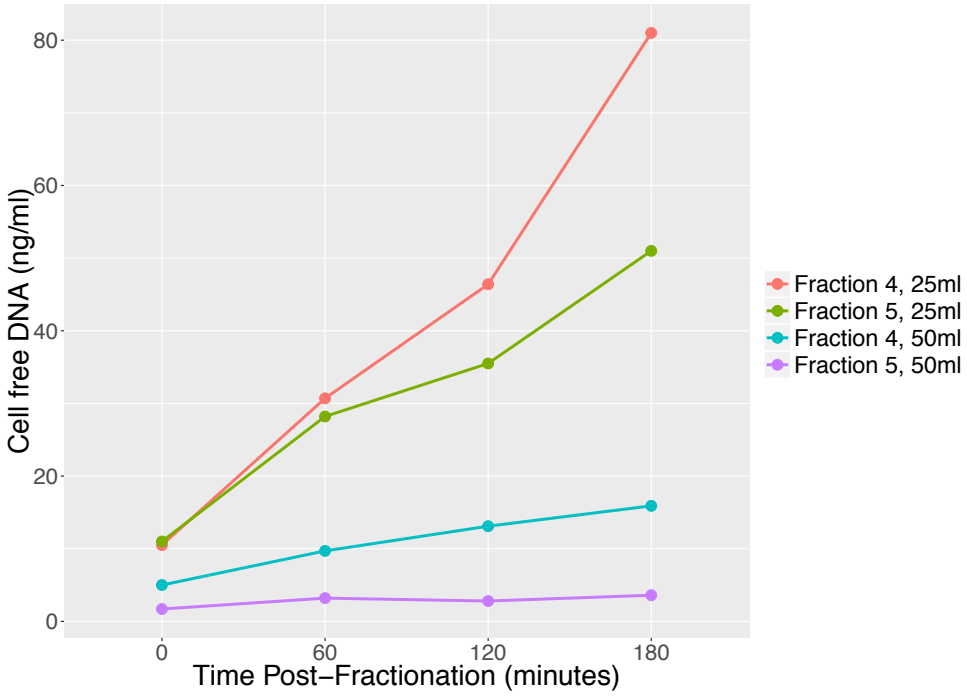

Fig. 3.3 Plot showing the concentration of cfDNA in different fractions of single crypts and showing the effects of varying the volume of PBS used for re-suspension. As expected, re-suspending in 50ml of PBS reduced the amount of cfDNA observed. Furthermore, fraction 5 appears to contain less cfDNA when compared with fraction 4.

In addition to sequence complexity, the length of the flanking region is also important. This is because each amplicon is ideally suited for paired-end 150bp sequencing by allowing a maximum of 70bp prior to the tract plus 60bp within the wild-type microsatellite therefore only 130bp reads are actually required giving a 20bp buffer at the 3'-end of the read. Furthermore, by having each amplicon length restricted to between 100bp (20bp at each flank plus 60bp tract) and 200bp (60bp at each flank plus 70bp tract), the length diversity within the library is restricted thus reducing any bias towards shorter amplicons. Before submitting for synthesis, each primer pair was manually curated for sequence diversity and primer pairs that produced amplicons at the extremes of the 100bp to 200bp range were avoided, where possible.

In addition to locus specific sequences, NGS adaptors were added to the 5' ends of both the forward and reverse primers to produce amplicons with sequencing adaptors at both the 5' and 3' ends. One of two sets of adaptors were used and are described fully in Section 2.10.2.

Primer pairs for 80 $[CA]_{30}$ microsatellites spanning all the chromsomes in the mouse genome were selected. Primer pairs were then tested for efficacy using a standard Phusion polymerase reaction in the TS_66_35 cycling protocol, Section 2.9. All primer pairs were tested in duplicate and assessed for the presence of a product at the expected length using gel electrophoresis. Out of 80 primer pairs, 63 primer pairs generated a product at the expected length (a 79% success rate). These 63 primer pairs were taken forward for further testing.

3.5.1 Preserving native $[CA]_{30}$ length during the Polymerase Chain Reaction

The first attempt at amplifying and sequencing $[CA]_{30}$ microsatellites from the mouse genome required 50 cycles of PCR and utilised the New England Biolabs Q5 DNA polymerase followed by Illumina HiSeq 4000 sequencing. This initial protocol lead to significant slippage of the microsatellite *in vitro* with the majority of reads in the $[CA]_{25}$ read bin instead of the $[CA]_{30}$ bin. Furthermore, the blurring effect was significant with many reads present at lengths different to that of the modal tract length, Figure 3.4A. It was obvious from this initial work that significant optimisation was required to maintain native tract length during amplification and sequencing.

In contrast, the final protocol required only 35 cycles of PCR, utilised the New England Biolabs Phusion DNA polymerase and could be sequenced using either the Illumina MiSeq

or HiSeq sequencing platforms. The outcome of this optimisation can be seen in Figure 3.4B. Below is a description of the key optimisation steps used to produce this sequencing data.

3.5.2 Selection of DNA polymerase for microsatellite amplification

The ideal DNA polymerase would have perfect fidelity, would never lead to an incomplete amplification reaction and would capture every amplicon during each cycle of the reaction. For this project, high efficiency and fidelity in amplification are key. By first identifying the polymerases with the highest efficiency i.e. the highest yield of product, the fidelity of these enzymes can be subsequently tested by sequencing of their products.

The majority of polymerases tested were thermostable polymerases (used in thermo-cycling reactions) but two thermolabile polymerases, that utilise isothermal amplification, were also tested. A range of polymerases were selected with varying characteristics that included 3'-to-5' exonuclease activity and the addition of an Sso7d domain that increases the polymerase-DNA interaction area.

The selected polymerases were trialled using the reaction conditions recommended by the manufacturer with an input of mouse reference DNA equivalent to the content of a single mouse crypt (1.5ng). Each reaction underwent 35 cycles of PCR or, in the case of the isothermal reaction, 90 minutes of incubation for the IsoAmp kit or 40 minutes for the TwistDx kit. The full list of polymerases trialled including key characteristics and results of gel electrophoresis assessment of are shown in Table 3.3.

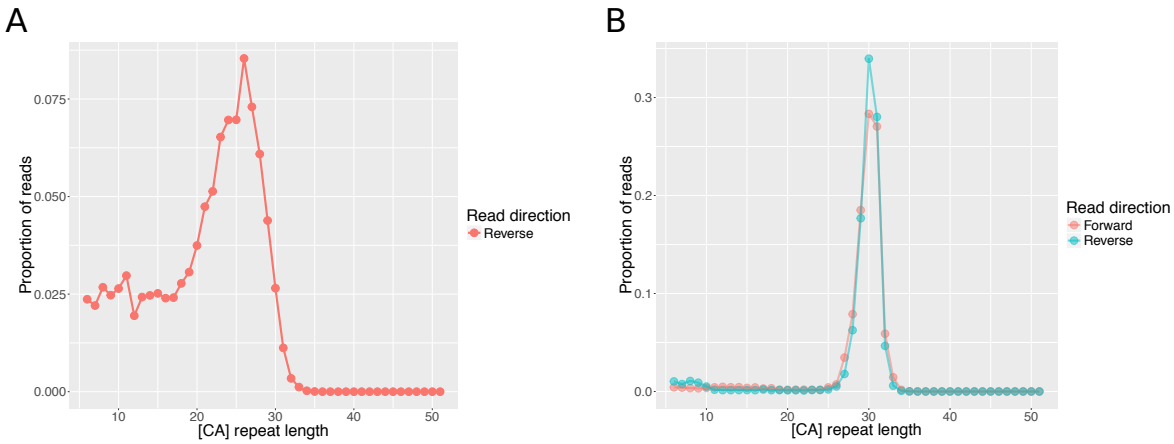

Fig. 3.4 (A) Scatter plot with line annotation showing initial attempts at trying to amplify an endogenous $[CA]_{30}$ microsatellite with the majority of reads slipping to $[CA]_{26}$ and a significant blurring effect present. (B) Scatter plot with line annotation showing the reduction in error as a result of PCR optimisation leading to the majority of reads being preserved at $[CA]_{30}$ with minimal distribution blurring.

Manufacturer	Polymerase	Type	Key features	Result
New England Biolabs	Q5	Thermostable	3'-to-5' exo, Sso7d domain	Moderate amount of product
New England Biolabs	Phusion	Thermostable	3'-to-5' exo, Sso7d domain	Good amount of product
New England Biolabs	Taq	Thermostable	5'-to-3' exo, 5'-flap endonuclease	No product detected
New England Biolabs	Vent	Thermostable	3'-to-5' exo	No product detected
Kapa Biosystems	HiFi	Thermostable	Taq developed by mutagenesis screen for amplicon balance	No product detected
TwistDx	TwistAmp Basic	Thermolabile	Recombinase polymerase technology, isothermal amplification at 37 °C	No product detected
BioHelix	IsoAmp II Universtal tHDA	Thermolabile	Helicase-dependent amplification, isothermal amplification at 65 °C	No product detected

Table 3.3 Table of DNA polymerases tested for $[CA]_{30}$ amplification either using 35 cycles of thermocycling PCR or, for the isothermal amplification, 90 minutes for the IsoAmp kit or 40 minutes for the TwistDx kit. All reaction were assessed by gel electrophoresis for presence of a product. Amount of product was assessed qualitatively by visual inspection of band brightness.

An intriguing observation from the testing of various polymerases was the improved product formation in the presence of the Sso7d domain. The Sso7d domain is a protein domain that was discovered in *Sulfolobus solfataricus* and has been shown to be highly stable at a high temperatures [38]. This domain has been added to many new generation polymerases as it increases the enzyme's interaction with target DNA from approximately 7-9 nucleotides in standard Taq to 12-14 nucleotides thus improving processivity [31, 88]. A potential consequence of this is that the polymerase is able to deal more readily with highly repetitive DNA sequences. At the very least, this data suggests the presence of the Sso7d domain improves amplification efficiency of microsatellites from low template copies.

From this screen, it was decided that the fidelity of the Q5 and Phusion enzymes (both manufactured by New England Biolabs) would be tested. Libraries were prepared using supra-optimal amounts of DNA, 10ng of input DNA (equivalent to 6.7 mouse crypts) with singleplex reactions targeted at 7 different loci so as to observe the efficacy of these enzymes at different genomic loci. The effect of the number of PCR cycles and the annealing temperature of the reaction were tested. To vary the amount of PCR cycles, all samples were initally amplified with 20 cycles of PCR followed by either 5 barcoding cycles or 15 barcoding cycles totalling 25 or 35 cycle of PCR respectively. To test the effect of annealing temperature, samples were run on a temperature gradient ranging from 66.2 °C to 71 °C. This totalled 168 reactions:

- 7 different loci = 7 samples
- Phusion versus Q5 = 2 test conditions
- 25 cycles versus 35 cycles = 2 test conditions
- 6 different annealing temperatures = 6 test conditions
- 7 samples x 2 enzymes x 2 cycling totals x 6 annealing temperatures = 168 reactions

The Q5 and Phusion polymerases both show significant improvement on previous attempts to amplify $[CA]_{30}$ microsatellites and both performed comparably in terms of read length distribution and consistency of error, Figure 3.5. Overall, both enzymes could theoretically be used for genotyping of wild-type loci. However, the Q5 enzyme was not able to generate adequate read depth for 3 of the 7 loci tested whilst the Phusion enzyme had a 100% success rate, Figure 3.5. This is further exemplified when the read depth spread of

these two polymerases is assessed, Figure 3.6. The Phusion enzyme produces significantly more reads per sample than the Q5 polymerase indicating superior amplification efficiency.

Annealing temperature was shown to have little effect on the performance of the Phusion polymerase, Figure 3.7. The lowest annealing temperature (66 °C) was used for all subsequent reactions so as to reduce the likelihood of extension reactions during the annealing step.

Finally, an assessment of the effect of PCR cycles was done comparing the samples that underwent 25 cycles of PCR compared with 35 cycles of PCR. The results of this are shown in Figure 3.8. As can be seen from the read length distributions, the samples undergoing 25 cycles of PCR have a modestly tighter read length distributions, as would be expected, and show that the final 10 cycles of PCR do have a slight blurring effect on the overall read length distribution. However, the number of reads per sample produced from 25 cycles of PCR is significantly lower than compared with 35 cycles of PCR, Figure 3.9, and for 3 of the 7 loci tested, there were not enough reads to generate an adequate read length distribution. The read depth is so low that generating a library with samples produced from 25 cycles of PCR alone is highly unlikely to be of high enough concentration to be suitable for Illumina sequencing. This issue could potentially be worked around by spiking in extra DNA to increase the overall library concentration. However, this would lead to significant data loss and be a highly inefficient method for generating crypt sequencing data. For this reason, the decision to use 35 cycles of PCR for future reactions was made.

From this data, the Phusion DNA polymerase can be seen to be the enzyme with the highest amplification efficiency and fidelity at repetitive stretches of DNA. Furthermore, an annealing temperature of 66 °C would be used with 20 cycles of initial PCR followed by 15 cycles of barcoding PCR.

3.6 Amplifying $[CA]_{30}$ loci directly from crypt lysate

The initial optimisation was done using DNA concentrations above that expected to be contained within a single crypt. To ensure that the protocol was also able to effectively amplify the genomic DNA content of a single murine crypt, amplification and sequencing of lysate containing a single murine crypt was performed. Furthermore, it was necessary to know if the presence of additional contaminants contained within crypt lysate would inhibit PCR. 46 primer pairs were selected from the successful panel of 63 primer pairs and amplified

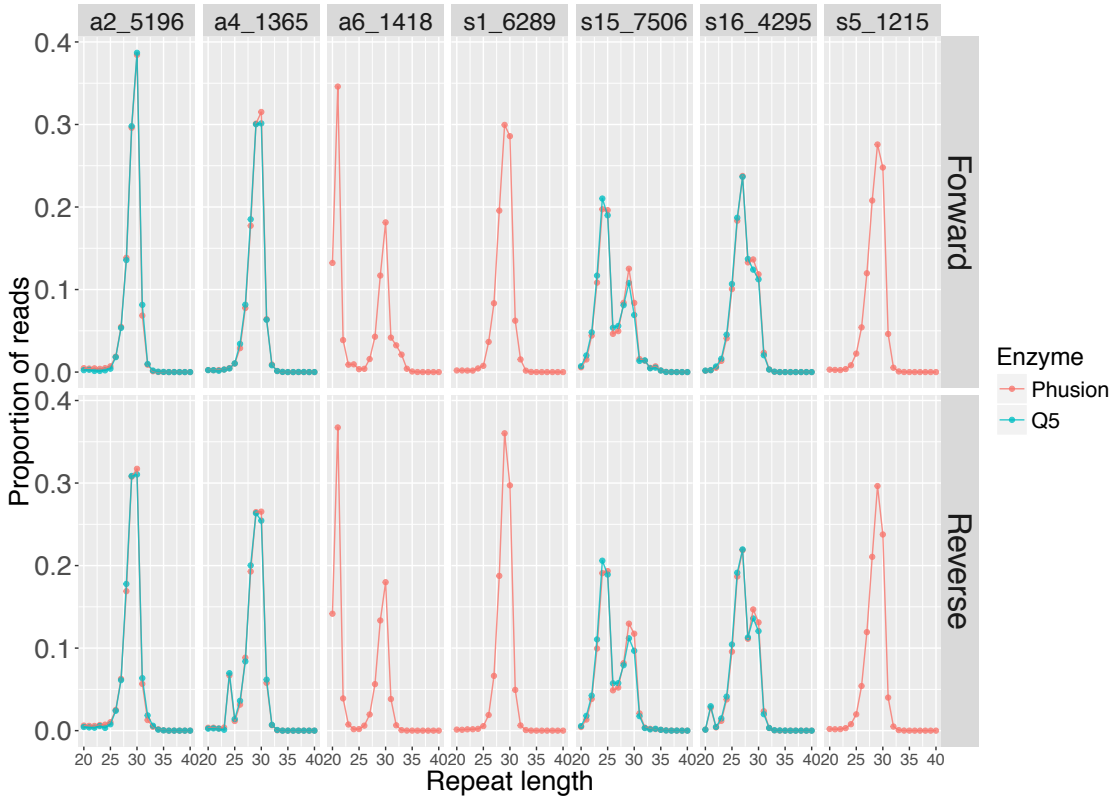

Fig. 3.5 Scatter plot with line annotation showing the read length distributions of the Q5 polymerase versus the Phusion polymerase at 7 different mouse loci. The distributions vary between different loci and between the forward and reverse reads generated by paired-end sequencing. The enzymes perform comparably in terms of read length distribution consistency and 'peakiness'. However, the Q5 polymerase did not produce adequate read depth at loci a6_1418, s1_6289 and s5_1215 suggesting poor amplification efficiency at these loci.

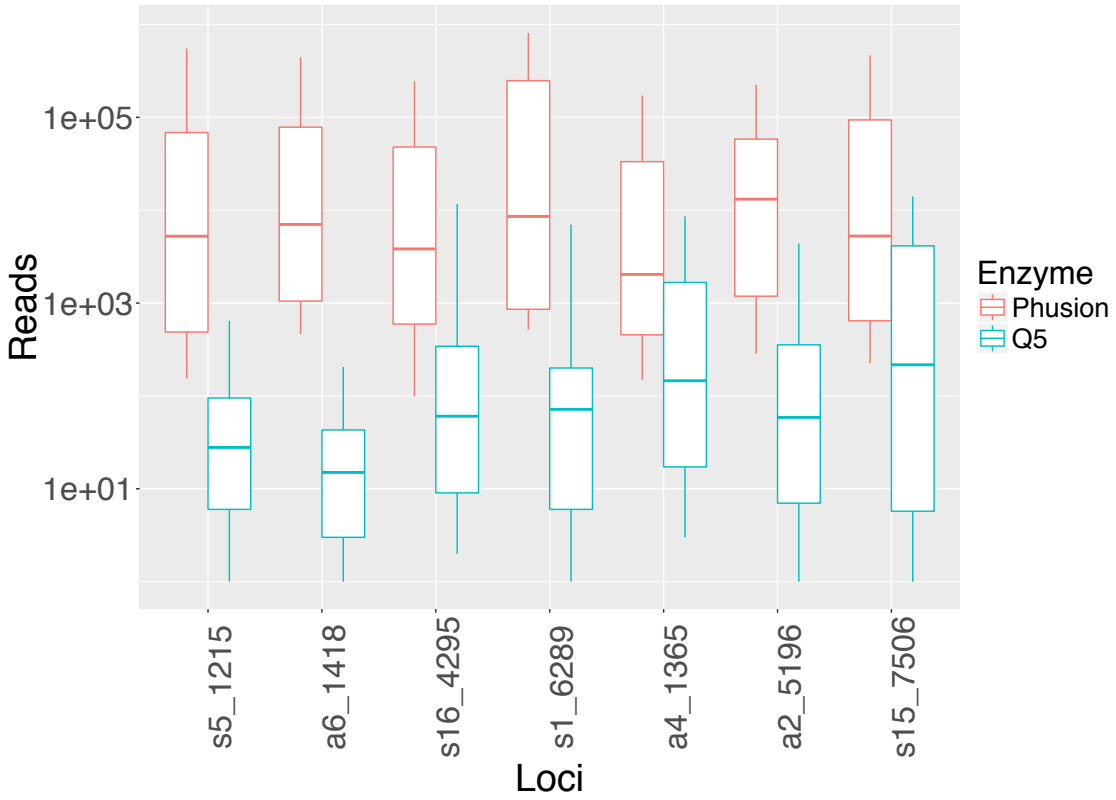

Fig. 3.6 Boxplot depicting the median read depth along with upper and lower quartiles and lowest and highest values across the 7 tested loci using either Phusion polymerase or Q5 polymerase for the library preparation. The Phusion polymerase produces substantially more reads across all loci indicating superior amplification efficiency.

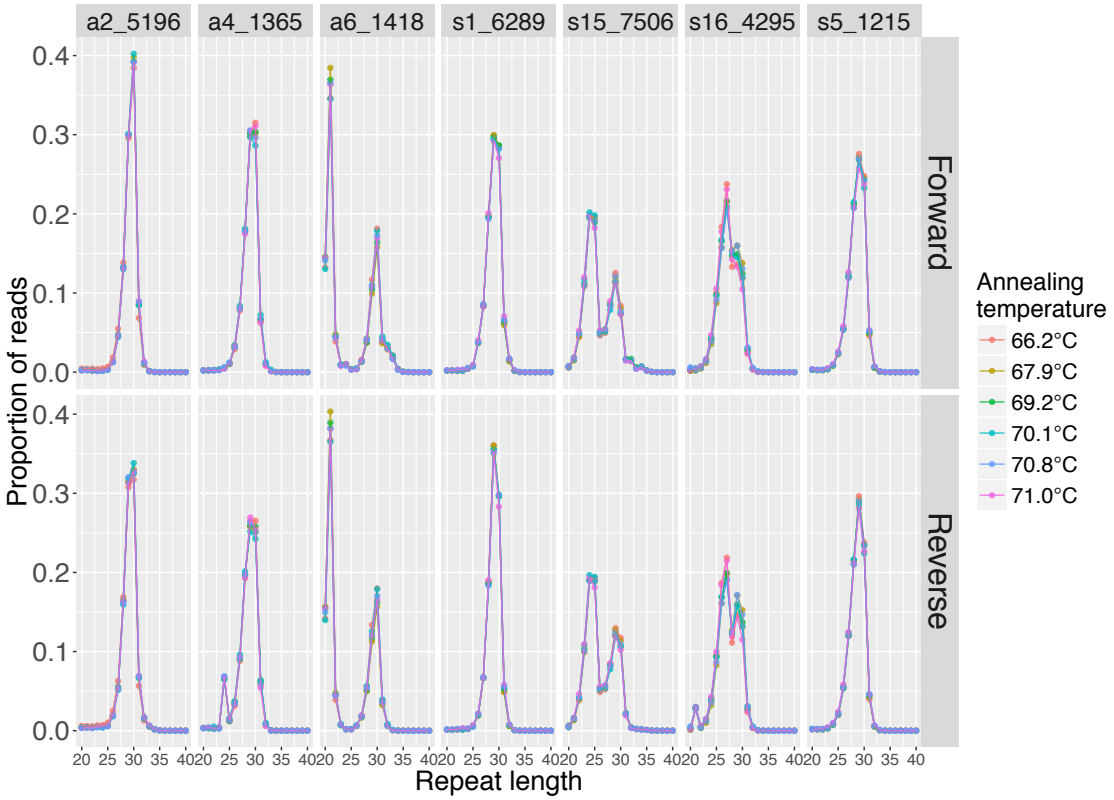

Fig. 3.7 Scatter plot depicting read length distributions at different loci whereby technical repeats were amplified at differing annealing temperatures. The read length distribution does not differ significantly between annealing temperatures.

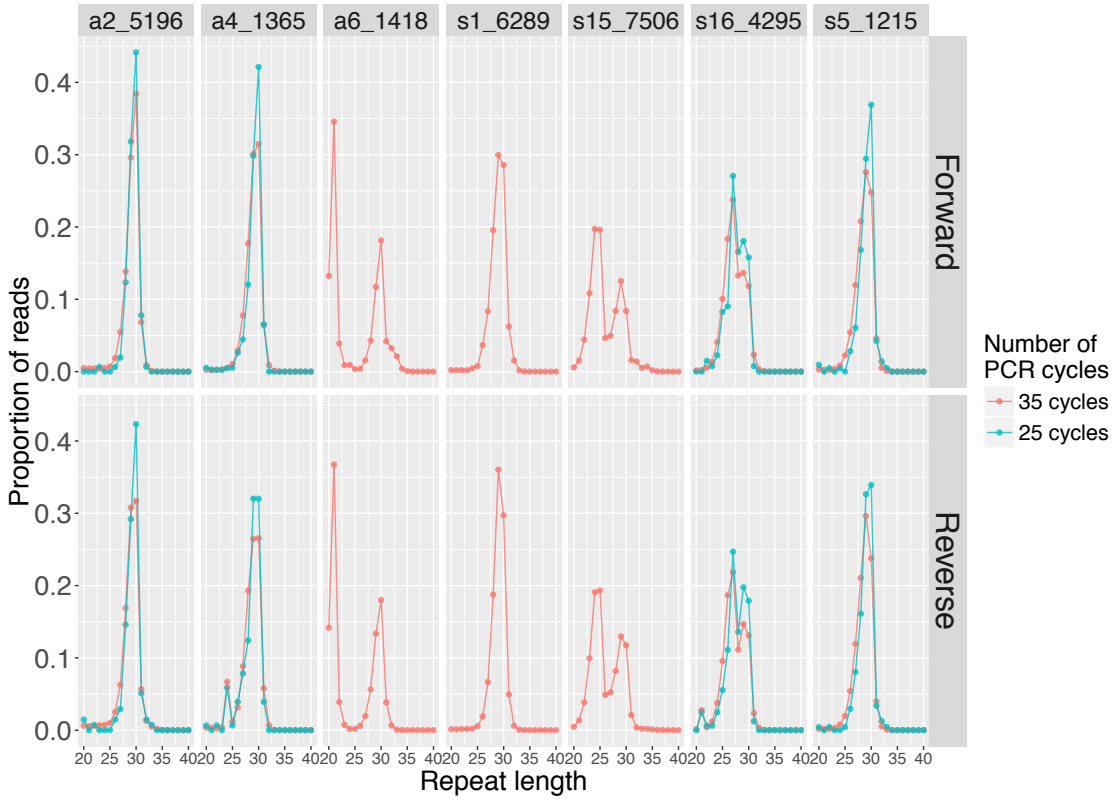

Fig. 3.8 Scatter plot showing the read length distributions generated from samples undergoing differing amounts of PCR cycles. The read length distributions of the 25 cycle samples are generally tighter indicating a small effect of an additonal 10 cycles of PCR on the blurring effect of the read length distributions. Dropout of reads at certain loci are due to read depth being too low to generate a read length distribution.

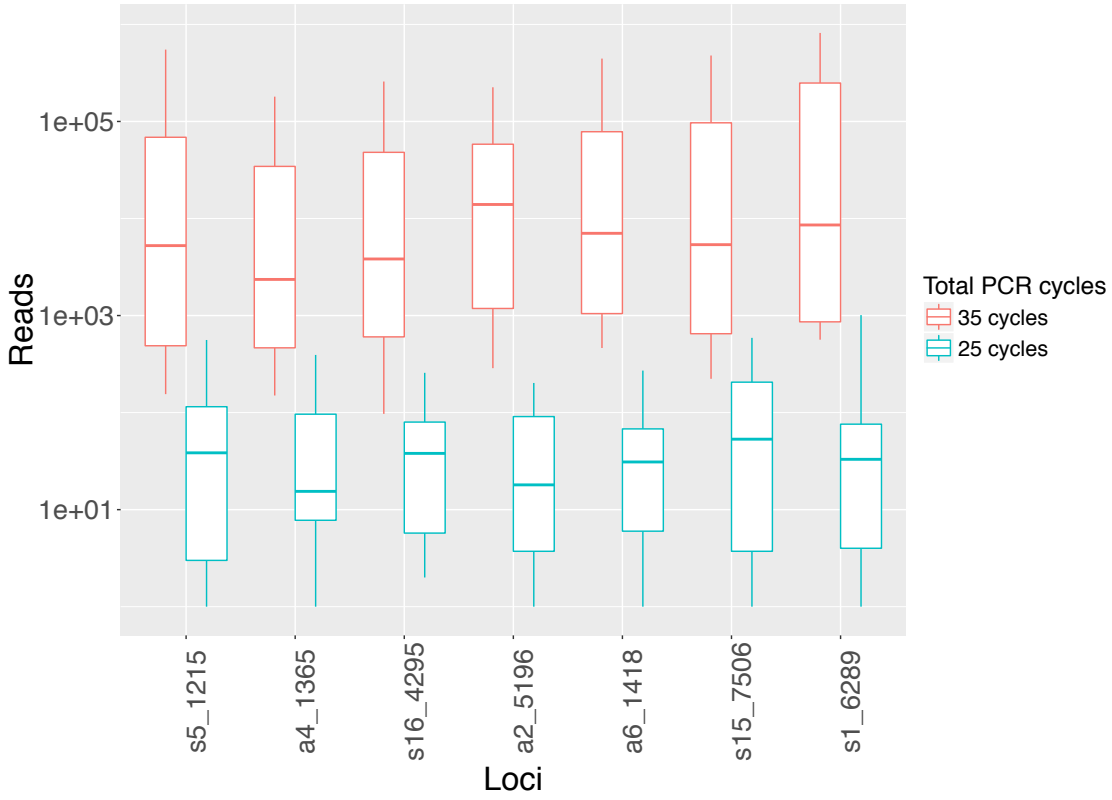

Fig. 3.9 Boxplot depicting the median read depth along with upper and lower quartiles and lowest and highest values across 7 different loci. Significantly fewer reads per sample are observed in samples undergoing 25 cycles of PCR compared with 35 cycles of PCR. This significant reduction in read depth will limit the feasibility of using low amounts of PCR cycles to generate sequencing data from single crypts.

92 crypts in singleplex, with each primer pair in duplicate. The products were assessed using gel electrophoresis showing 45 out of the 46 primer pairs had successfully amplified a discrete band at the appropriate size for the locus.

To ascertain any effect of the contents of the crypt lysate on read length distribution, the same primers applied previously to amplify crypt equivalents were used to amplify lysate containing a single crypt. As can be seen in Figure 3.10, the read length distributions for both the crypt equivalents and real crypts had a high degree of similarity. However, there are some subtle differences between the read length distributions generated. The possible sources of this are: a) additional contaminants present in the crypt lysate effecting polymerase fidelity, b) batch effects as a result of the crypts being run on a separate lane of sequencing to the crypt equivalents or c) normal error seen between technical replicates. The effect of additional contaminants could be overcome by optimisation of the PCR conditions and batch effects can be controlled for by supplying separate reference samples for each sequencing run to allow the building of reference distributions unique to each sequencing lane. However, a concern is that this error is a normal consequence of stochastic polymerase error such that a crypt distribution may differ from its reference distribution based upon stochastic effects alone. This is tested further in Section 3.8.

In addition to comparing the read length distributions generated by crypt equivalent samples compared with single crypt lysate, we hypothesised that microsatellites displaying length variation in the germline may have higher somatic mutation rates. Reference samples from 6 different mice were isolated and 63 loci were amplified from each sample. Out of the 63 loci studied, 58 loci were called as having the same length in all 6 mice. However, 5 loci showed variation between the 6 individuals, Figure 3.11. It is likely that the difference in microsatellite length at these loci is due to germline mutations however, the possibility of crossing with a contaminating polymorphic strain cannot be ruled out. Wherever possible, these loci were included in future analyses to see if an increase in somatic mutability was observed.

3.7 Estimating minimal read depth requirements

An important metric to ascertain from sequencing of microsatellites is the effect of read depth on signal-to-noise ratios. The signal-to-noise ratio can be assessed by taking samples that have a length call of $[CA]_{30}$ and dividing the number of reads at $[CA]_{31}$ by the number

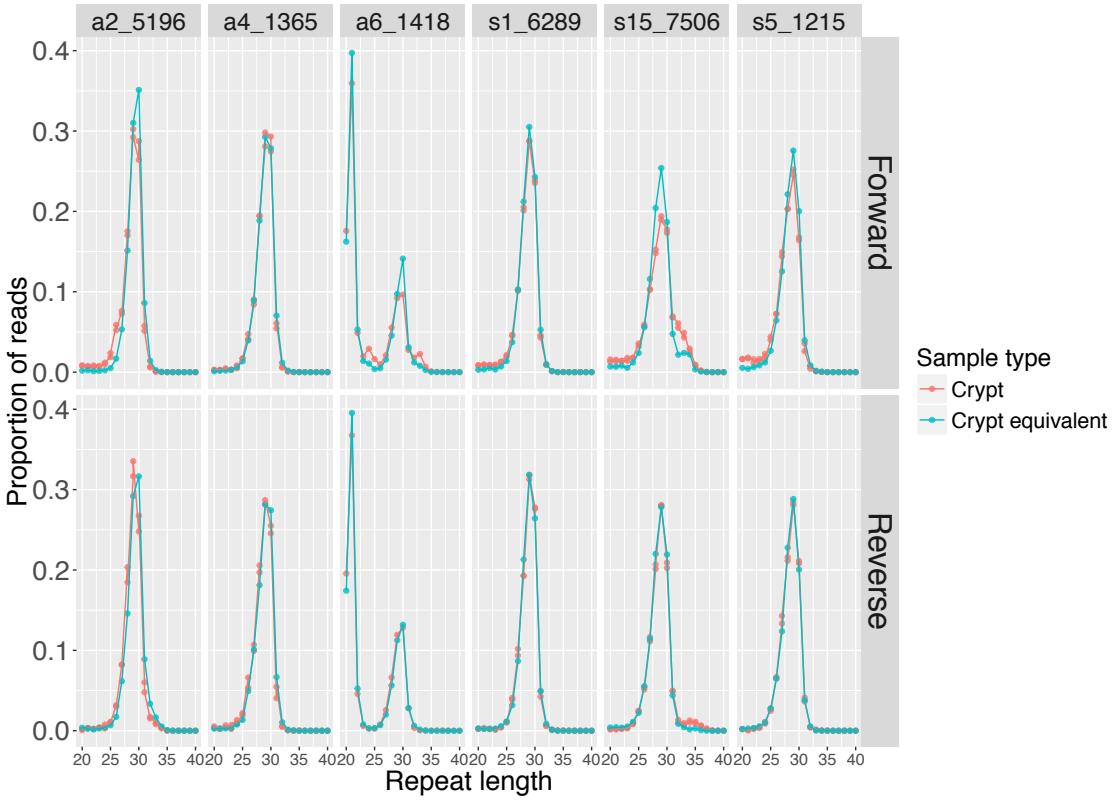

Fig. 3.10 Scatter plot with line annotation comparing the read length distributions generated from crypt equivalent samples versus real crypt samples. The read length distributions in the real and crypt equivalents appear very similar.

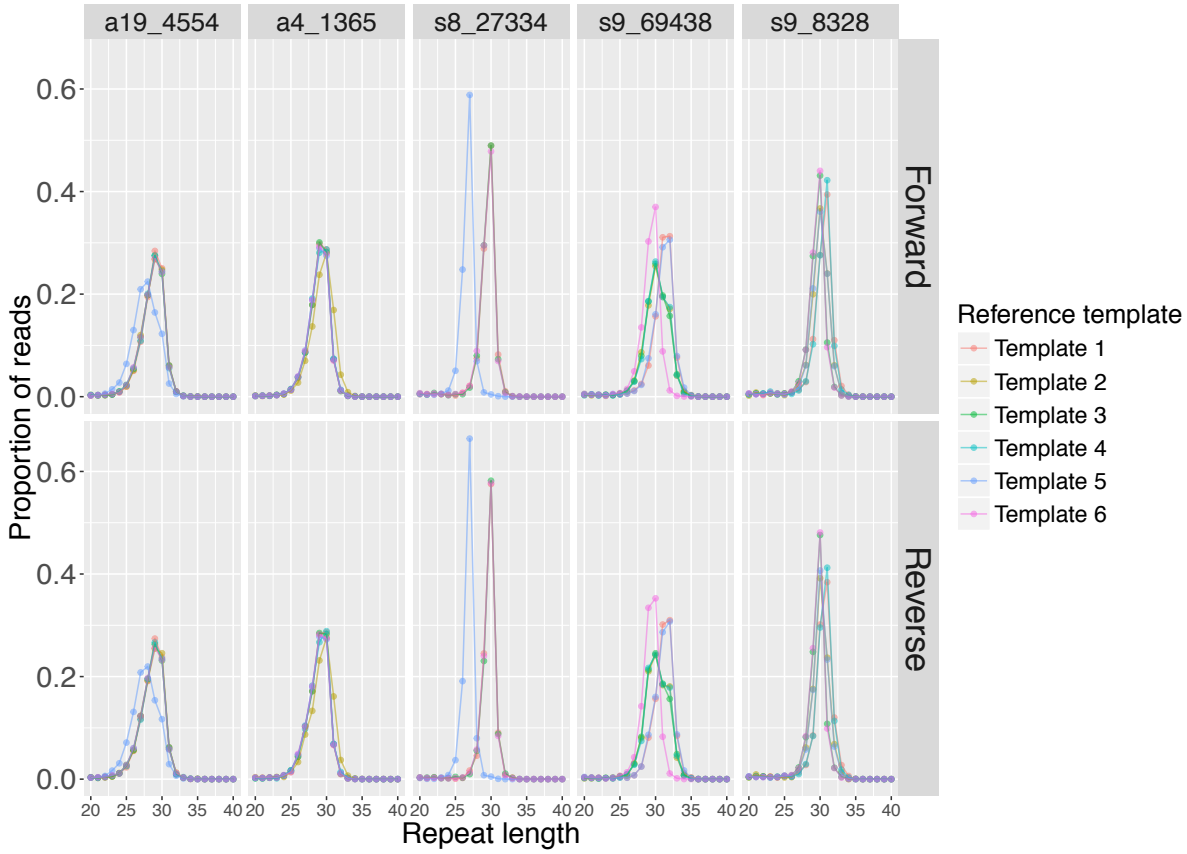

Fig. 3.11 Scatter plot with line annotation showing the read length distributions of reference templates from 6 different mice. 5 of the loci show differing length calls across the individuals and were hypothesised to be more somatically mutable.

of reads present at $[CA]_{30}$. By plotting this metric against the total number of reads within the distribution (i.e. between $[CA]_{20}$ and $[CA]_{35}$), the lower limit of read depth can be determined. As can be seen in Figure 3.12, the lower limit of read depth appears to be 1000 reads within the distribution. Therefore, to ensure consistent signal-to-noise, a filter was added to the analysis pipeline to remove all loci with less than 1000 reads within the distribution.

3.8 Testing consistency of polymerase error

A high *in vitro* polymerase error is an inevitable consequence of amplifying microsatellites without the presence of mismatch repair mechanisms. In order to accurately quantify the mixture of wild-type and mutant microsatellite lengths within a crypt lysate, the level of variance in the error made by the polymerase must be consistent between technical replicates. Minimal variance between technical replicates will enable confident separatation of biological variance (i.e. the presence of a mutant clone) from technical noise. To test the level of variance between technical replicates, 15 loci were amplified from 1.5ng of DNA (equivalent to one murine crypt) in 8 technical replicates. The read length distributions show remarkable similarity at all loci in both read directions, Figure 3.13, showing that the level of technical noise is very low and that detecting biological deviance from these distributions should be possible.

An intriguing observation from this test was the loci and read direction specific nature of the polymerase error. Each loci appears to have a different error distribution and is likely due to a combination of polymerase error during PCR and sequencing error on the Illumina platform. The only variable between the different loci is the flanking region sequence and thus must be the determinant of polymerase and sequencing error. The exact property of the flanking sequence that determines the level of error remains unknown. Furthermore, the error distributions between the forward and reverse reads at a given locus also differs significantly. A key example of this is the forward and reverse read distribution differences seen at locus s15_7506 in Figure 3.13. This distribution is so significantly different that the loci could not be used in any downstream analysis. As the forward and reverse reads are generated from the same molecule, the only determinant of the difference in forward and reverse read length distributions must be sequencing error. Whether this is a consequence of sequencing chemistry infidelity or a loop-hole in the sequencer's read building algorithms

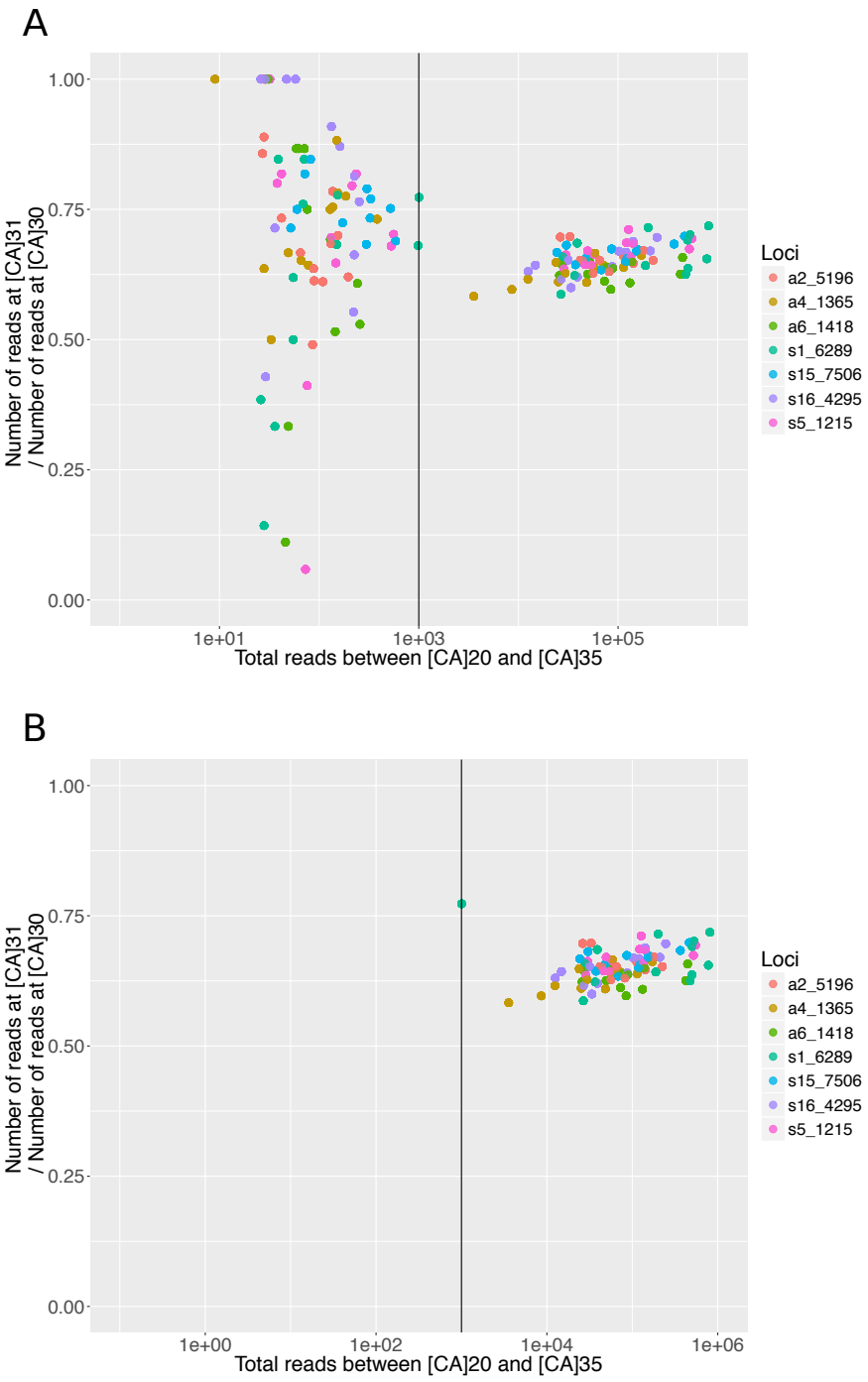

Fig. 3.12 Scatter plot showing the ratio of reads at $[CA]_{31}$ to reads at $[CA]_{30}$ in samples with a length call of $[CA]_{30}$ as a function of total reads between $[CA]_{20}$ and $[CA]_{35}$. The vertical line intercepts the x-axis at 1000 reads. (A) All samples shown with a wide range of signal to noise at low read depth. (B) All samples with less than 1000 reads are removed, signal to noise is far more stable.

is unknown. Regardless, the key finding is that the read length distributions, and therefore the level of error, is consistent between technical replicates and should allow detection of biological differences. Furthermore, the need for analysing each loci and each read direction separately is highlighted.

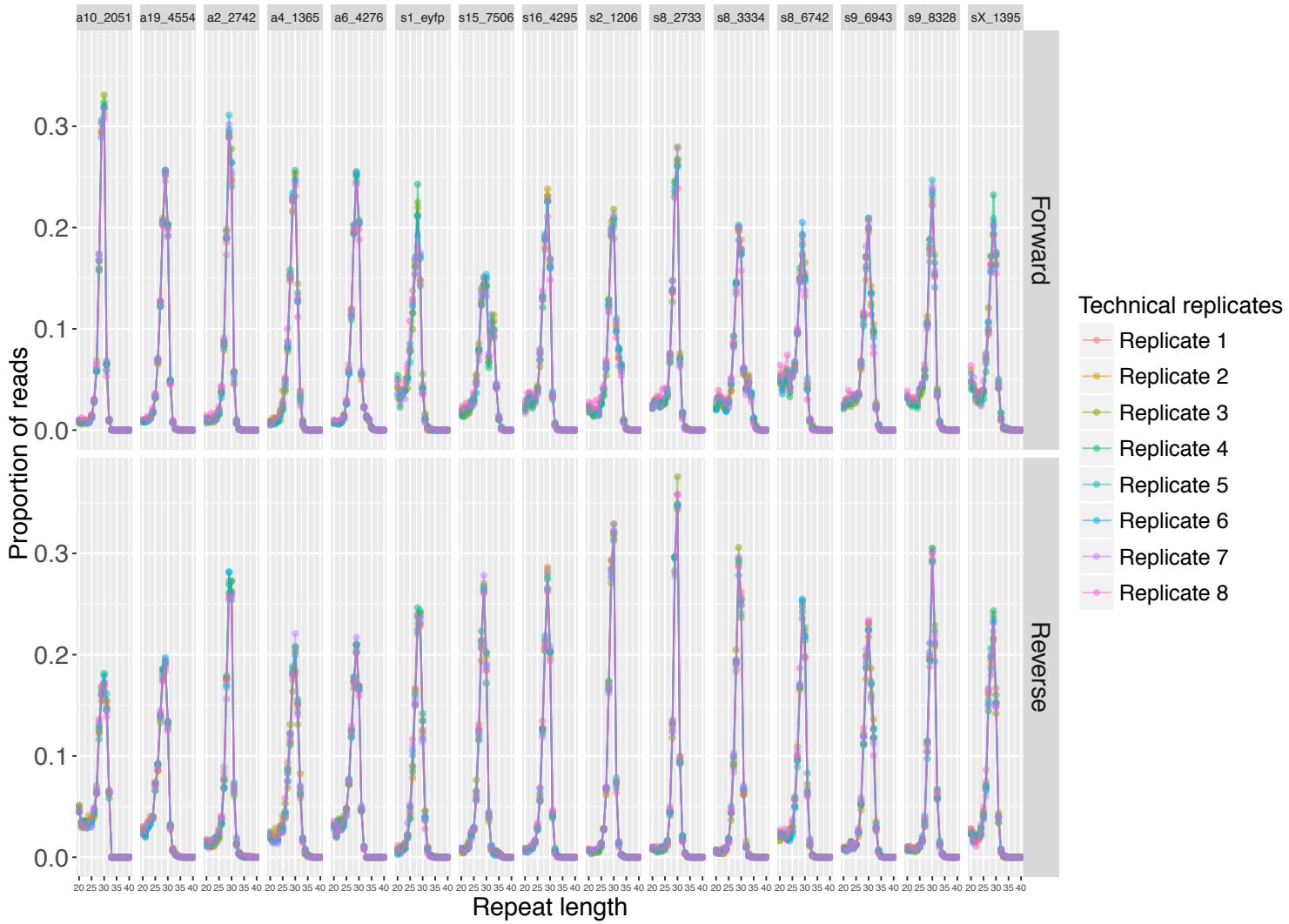

Fig. 3.13 Scatter plot with line annotation showing read length distributions at 15 different loci in the forward and reverse direction. Each distribution was generated from 1.5ng of mouse DNA with 8 technical replicates for each. The read length distributions show remarkable consistency in error across all 15 loci in both read directions.

	MiSeq	**HiSeq 4000**
Maximum read length	2x300bp	2x150bp
Maximum reads per lane	25M	625M
Cost per megabase (based on 2x150bp)	8.8p	0.7p

Table 3.4 A comparison of the key parameters between the Illumina MiSeq and HiSeq 4000 sequencing platforms.

3.9 Comparison of sequencing error on the MiSeq platform compared with the HiSeq 4000 platform

Illumina have a large repertoire of sequencing platforms commercially available. The key difference between each platform is the number of reads generated per lane, the length of read available and the associated per base cost of sequencing. Initially, the Illumina MiSeq platform was used for optimisation of the protocol. This was due to its rapid sequencing time and amenability to high read depth sequencing of amplicon libraries. In later experiments, the HiSeq 4000 platform was used. This platform generates a large amount of sequencing data with 8 lanes per flow cell and a slightly longer run time. Prior to the HiSeq 4000, the longest read length available for high yield sequencing was 125bp. Fortunately, with the introduction of the HiSeq 4000, read lengths of up to 150bp became available allowing libraries generated from the same primer pairs used for MiSeq library preparation to be loaded. The key features of the MiSeq platform compared with the HiSeq 4000 are summarised in Table 3.4.

As all of the optimisation data was generated on the MiSeq platform, it was important to test the consistency in read length distribution when the same library was sequenced on the MiSeq and HiSeq 4000. As the HiSeq 4000 generates approximately 10-fold more reads than the MiSeq, the use of the HiSeq will ensure maximal read depth for all loci in all crypts sampled whilst also driving down the cost of analysing each crypt. As can be seen in Figure 3.14, the read length distributions are very similar when sequencing the same library on the HiSeq when compared with the MiSeq.

In addition, there is a slight improvement in the tightness of the peaks generated when sequencing on the HiSeq as well as greater consistency between samples (for example, locus s8_2733, forward read). It is likely that the improved consistency is as a result of

Locus	Lengths isolated
a4_1365	$[CA]_{27}$, $[CA]_{28}$, $[CA]_{29}$, $[CA]_{30}$
s9_8328	$[CA]_{30}$, $[CA]_{31}$, $[CA]_{33}$

Table 3.5 Summary of synthetic loci generated from the plasmid cloning protocol outlined in Section 3.10 and depicted in Figure 3.15

markedly increased sequencing depth on the HiSeq (mean number of reads within each locus distribution on the HiSeq and MiSeq was 29,050 and 7,291 respectively). However, it seems unlikely that the increased read depth will be contributing to the improved tightness of the distributions observed at some loci on the HiSeq. This improvement in read length distribution is likely to be caused by reduced sequencing error on the HiSeq when compared with the MiSeq revealing that some of the read length distribution blurring is caused by the sequencer and is not due to the initial PCR polymerase error alone.

3.10 Validation of PCR method using synthetic loci

A potential caveat of using mutant microsatellites as clonal marks is the known bias in PCR amplification for shorter molecules. A consequence of this is that a mutant microsatellite, e.g. $[CA]_{28}$, would be over amplified with respect to the wild-type $[CA]_{30}$ microsatellite leading to an over-estimate of the mutant clone size. A method for dealing with this is to only use small changes in microsatellite length as clonal marks thus reducing the probability of amplification bias. As an added control measure, two synthetic loci were generated with varying $[CA]_n$ lengths so as to assess any amplification bias using qPCR. To do this, two endogenous $[CA]_{30}$ loci were amplified using PCR to obtain a range of $[CA]_n$ repeat lengths. The pool of $[CA]_n$ molecules were then ligated into a pUC-19 plasmid and transformed into *E. coli*. This process is depicted in Figure 3.15 and described in Section 2.11. In total, 7 different lengths of microsatellite were generated representing two different genomic contexts, these stocks of synthetic loci are summarised in Table 3.5. Maxi preps of plasmid containing microsatellites of the desired length were validated by Sanger sequencing.

3.10.1 Synthetic loci distributions are comparable to endogenous loci

Before using the synthetic loci further, it is important to observe that the synthetic loci generate comparable distributions following amplification and sequencing to loci amplified

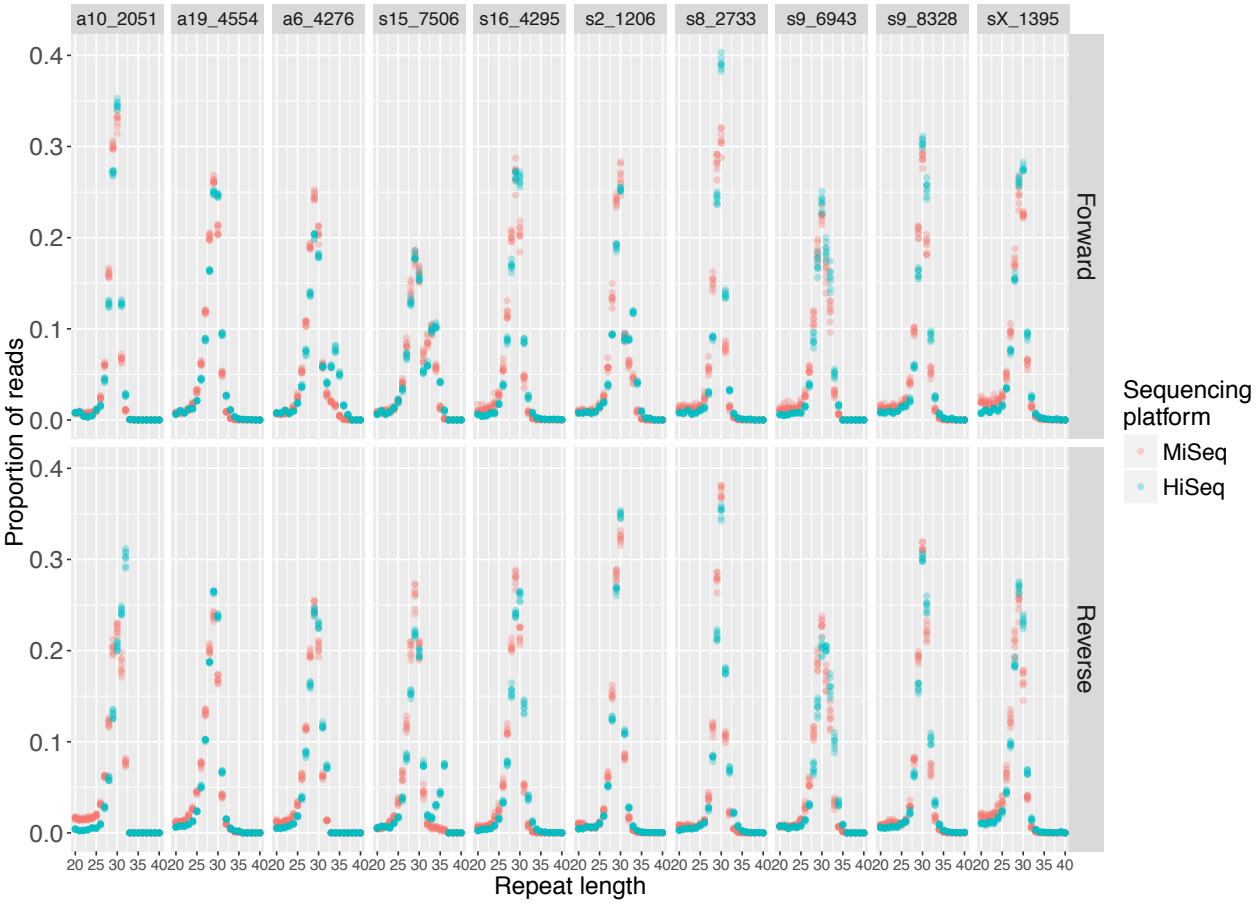

Fig. 3.14 A scatter plot comparing the read length distributions generated from the Illumina HiSeq 4000 and MiSeq sequencing platforms. The HiSeq 4000 displays improved consistency of read length distributions between technical replicates and tighter read length distributions at some loci. The improved consistency is likely to be as a result of increased read depth generated on the HiSeq (mean number of reads within the read length distribution on the HiSeq 4000 and MiSeq was 29,050 and 7,291 respectively). It is unlikely that the tighter read length distributions are caused by improved read depth and is likely to be as a result of higher fidelity of sequencing microsatellites using the HiSeq 4000 sequencing chemistry.

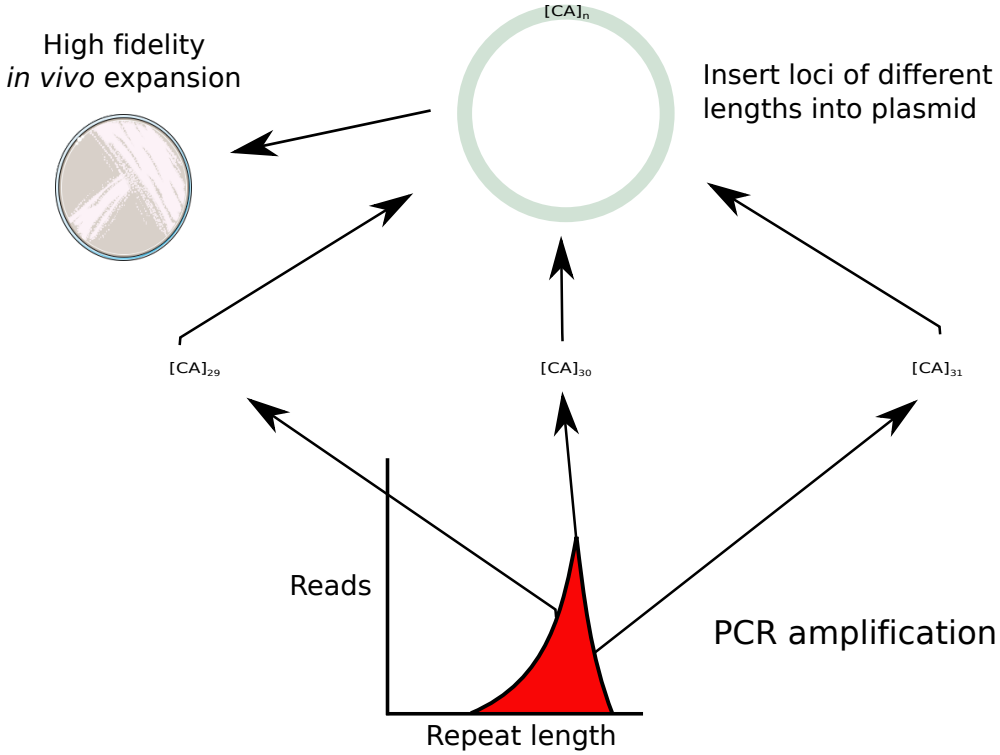

Fig. 3.15 PCR was used to generate a range of $[CA]_n$ loci before using high fidelity bacterial cloning to expand loci of varying lengths.

directly from genomic DNA. This was done by linearising and accurately quantifying the concentration of pUC-19 plasmid before diluting and adding 500 plasmid copies (equivalent to the number of genome copies per murine crypt) to a PCR reaction. These plasmid copies were put through the identical amplification and sequencing protocol to that used for amplifying crypt lysate. The resulting comparison of distributions from plasmid and genomic DNA can be seen in Figure 3.16. For locus a4_1365, the distributions are near identical. Surprisingly for locus s9_8328 the plasmid produced a far tighter distribution when compared to the same loci being amplified and sequenced from its genomic context. The cause of this difference is unknown but could be due to the chromatin context of the loci reducing polymerase fidelity in the early stages of amplification or, less likely, is the possibility that a single nucleotide polymorphism is present in the genomic DNA leading to reduced primer efficiency at this locus when compared with the plasmid DNA. Regardless of the cause of the difference, the results from locus s9_8328 should be approached with some caution whilst locus a4_1365 in the plasmid appears to behave in a similar manner to its genomic counterpart.

3.10.2 qPCR comparison of synthetic loci at different lengths reveals no PCR amplification bias

PCR bias towards smaller amplicons is well documented. If changes in the length of the $[CA]_n$ microsatellite were to lead to bias toward the shorter of the microsatellite amplicons, this would lead to the shorter amplicon being overrepresented and an over-estimate of the size of that clone. To quantify any PCR amplification bias, real-time quantitative PCR (RT-qPCR) was performed, as described in Section 2.16.2, on the 7 different lengths of $[CA]_n$ microsatellite representing two different loci. The result of this analysis can be seen in Figures 3.17 and 3.18.

As can be seen, in Figures 3.17 and 3.18, there appears to be no PCR amplification bias present between microsatellites of varying lengths at locus a4_1365 nor at s9_8328. However, the key difference between this analysis and the amplification of loci from the single crypt lysate is the number of loci copies. As qPCR analysis requires high template input, the number of copies added to each qPCR reaction was approximately 10^6 fold higher than the content of a single murine crypt. It is therefore possible that when the number of loci copies is reduced further any amplification bias may be exaggerated, particularly during the early rounds of PCR. A way of addressing this question is to perform a mixing experiment

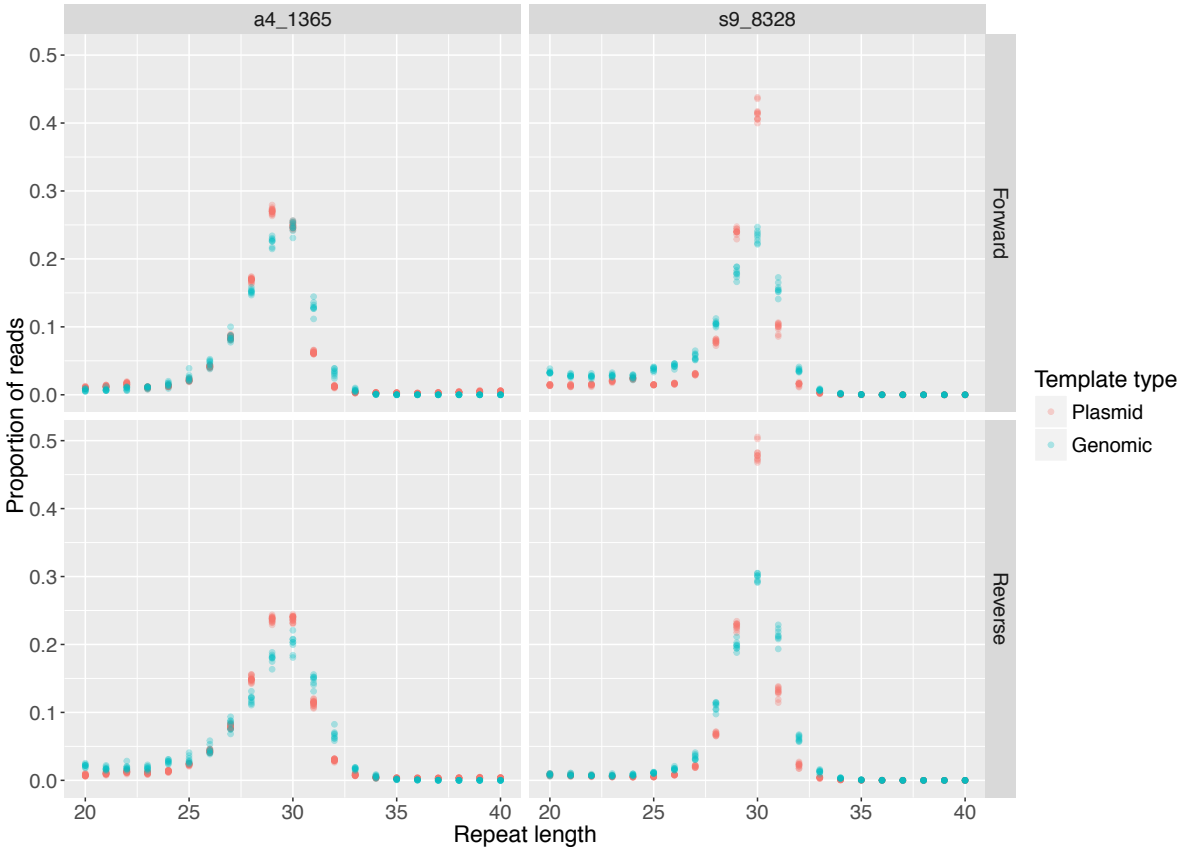

Fig. 3.16 Scatter plot showing the read length distribution following amplification and sequencing of the same $[CA]_{30}$ within a plasmid versus within genomic DNA.

to observe any bias in amplification when low amounts of template are amplified; this is discussed further in Section 4.3.

3.11 Multiplex PCR

As can be seen in Table 3.1, in order to adequately empower the study, over 8 microsatellites from a single crypt need to be amplified. The frequency of mutation represented in Table 3.1 is the frequency with which a single $[CA]_{30}$ tract mutates to bring the eYFP reporter in frame with small changes in length being far more likely than large alterations e.g. $[CA]_{30}$ to $[CA]_{31}$ being far more likely to occur than $[CA]_{30}$ to $[CA]_{13}$. However, the estimates from Kozar et al's data is based upon the observation of a single $[CA]_{30}$ loci present hemizygously. Using endogenous microsatellite sequencing as a clonal mark allows the use of biallelic loci and the detection of mutations that do not cause in-frame changes. The calculations shown in Table 3.1 take into account a doubling in mutation rate as a result of biallelic loci but do not take into account the possibility of detecting mutations that would not cause an in-frame shift. Overall, these estimates provide a conservative estimate of the expected number of WPCs and PPCs that will be observed using microsatellite sequencing.

Due to primer-primer interactions within the PCR reaction, the critical roadblock to effective multiplex PCR is the formation of primer-dimers. Due to their significantly shorter length when compared with the specific product size, the primer-dimer products are preferentially amplified during the PCR process. It is therefore necessary to optimise the multiplex PCR to increase specificity for the desired product. As these primer-dimers contain sequencing adaptors, once barcoded, these DNA fragments preferentially bind to the sequencing flow cell and significantly reduce the amount of sequencing data obtained from the flow cell. Furthermore, due to the relatively low complexity of primer-dimer molecules, their presence will reduce the quality of all sequencing being produced from the flow cell. Therefore, even small amounts of primer-dimer, generally >1% of the total molecules in the sequencing library, can significantly reduce sequencing data yield and overall sequencing quality. As a quality control measure, all libraries produced from multiplex PCR were checked for primer-dimer content using the Agilent Bioanalyser described in Section 2.16.1 and size selection performed using the optimised MagJET NGS size selection kit described in Section 2.15. Overall, an optimised multiplex PCR protocol with specific product amplification and reduced primer-dimer generation provides a good option for the sequencing of

multiple loci from a single crypt.

3.11.1 Optimisation of Phusion for multiplex PCR

This section describes the iterative series of experiments used to develop an optimised multiplex PCR protocol. Due to its superior performance in singleplex PCR, an initial attempt to optimise the Phusion DNA polymerase for use in a multiplex reaction was performed. The online tool, MultiPLX2.1 [52], was used to design optimal multiplex PCR groups using the primer pairs previously optimised for singleplex PCR. This yielded 6 different multiplex groups with 7 or 8 primer pairs per group which, when trialled on crypt equivalents containing 1.5ng of mouse DNA, produced adequate amounts of product but additionally produced large amounts of primer-dimer that required size selection prior to barcoding each sample. In addition, when trying to expand the multiplex groups beyond 8 primer pairs, large amounts of primer-dimer were produced preventing size selection entirely therefore making library preparation impossible. Therefore, a stepwise development process to optimise the Phusion DNA polymerase for the multiplex PCR of dozens of $[CA]_{30}$ microsatellites was performed. The guidance of Markoulatos et al [60] was invaluable in designing this development protocol. The steps taken in this optimisation are summarised in Table 3.6 and the final reaction mixture is described in Section 2.9.3.

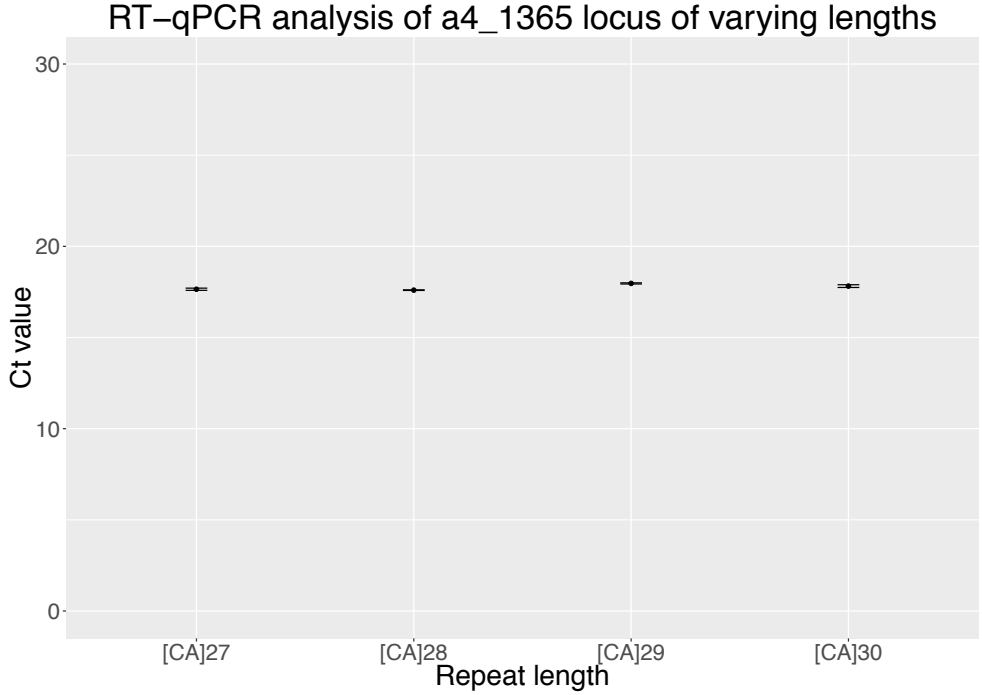

Fig. 3.17 Scatter plot showing the mean Ct value for differing microsatellite lengths at locus a4_1365 with error bars indicating standard deviation of replicates (N = 8).

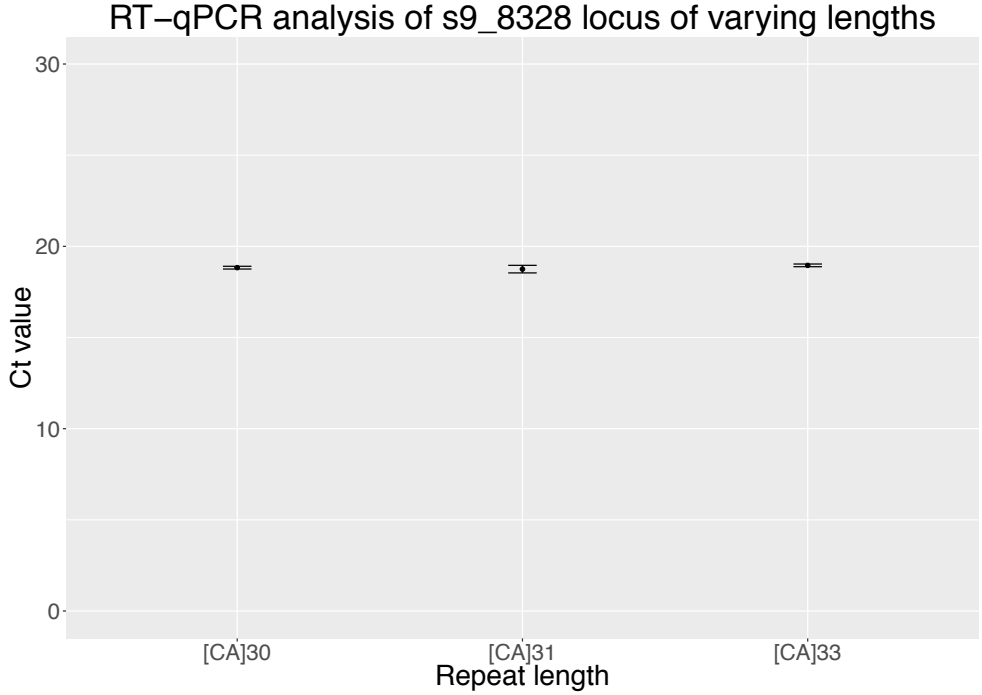

Fig. 3.18 Scatter plot showing the mean Ct value for differing microsatellite lengths at locus s9_8328 with error bars indicating standard deviation of replicates (N = 8).

Parameter tested	Conditions tested	Optimal result
Polymerase	Phusion vs Phusion HotStart	Phusion HotStart
Sequencing adaptor	CS1 vs M13	M13
Mg^{2+}:dNTP ratio	Mg^{2+}: 1.5mM, 2.0mM, 2.5mM, 3.0mM dNTP: 0.1mM, 0.2mM, 0.5mM, 1.0mM, 1.5mM	2.5mM Mg^{2+} : 1.0mM dNTP
Buffer concentration	1x, 1.5x, 2x, 2.5x	1x
Primer pool concentration (8 primer pairs)	4μM, 2μM, 1μM, 0.5μM, 0.3μM	2μM (0.25μM per primer pair)
Annealing temperature	56.0°C, 56.4°C, 57.4°C, 58.6°C, 60.5°C, 62.8°C, 65.5°C, 67.8°C, 69.5°C, 70.8°C, 71.7°C, 72.0°C	65.5°C
Polymerase amount	0.2U, 0.4U, 0.6U, 0.8U, 1.0U	0.8U
Additional glycerol	0%, 1%, 2%, 5%, 7.5%, 10%	0%
Betaine additive	0M, 0.5M, 1.0M, 1.5M, 2.0M	0M

Table 3.6 Table summarising the parameters that were varied to optimise the Phusion DNA polymerase for use in multiplex PCR. All of the above conditions were tested in the order that they are presented and were tested using a multiplex group containing 8 primer pairs. The final multiplex PCR conditions were able to simultaneously amplify up to 24 amplicons, a 3-fold increase when compared to multiplex group size before optimisation.

As a result of this optimisation process, it was found that the HotStart version of the Phusion DNA polymerase significantly reduced primer-dimer formation. Though the PCR reaction setup is done on ice, there is an inevitable basal level of DNA polymerase activity prior to thermocycling. The HotStart version of the polymerase utilises a haptamer-based enzyme activity blockade until the enzyme is heated beyond 55 °C. To ensure that the structural differences within the HotStart enzyme did not effect the ability of the polymerase to amplify $[CA]_{30}$ microsatellites, analysis of multiplex amplification of 6 technical repeats of murine crypt equivalents with 13 primer pairs was performed. This was compared with loci previously amplified using the standard polymerase but, due to the smaller multiplex group sizes in standard Phusion multiplexing, the loci were drawn from different multiplex groups. A comparison of the read length distributions generated by standard Phusion versus Phusion HotStart can be see in Figure 3.19. Both the standard Phusion and Phusion HotStart enzymes produced highly consistent read length distributions. The standard Phusion enzyme does produce slightly tighter read length distributions but the overall benefit of larger multiplex groups with the HotStart enzyme outweighs any benefit gained from tighter read length distributions using standard Phusion thus the HotStart enzyme was taken forward for future use.

As well as allowing larger multiplex groups and consistently performing well at microsatellite length calling, the Phusion HotStart enzyme also displays good amplicon balance within multiplex groups. The multiplex group chosen for future use was M13_33 as it displayed good read balance between 14 primer pairs (or 15 primer pairs when primers were added to amplify the $[CA]_{30}$ transgenic microsatellite). Each time the multiplex group was used, the concentration of each primer pair was adjusted to try and balance the read depth between amplicons within the group. The final concenrations of each primer pair is shown in Table A.3 and the resulting amplicon balance is compared with the initial amplicon balance in Figure 3.20. Further information about each of the loci amplified can be found in Appendix B.

3.12 Discussion

In this chapter, the stepwise development of a protocol for the multiplexed amplification and sequencing of $[CA]_{30}$ microsatellites from the genomic content of a single murine crypt has been described. This technique should allow for the accurate quantification of mutant

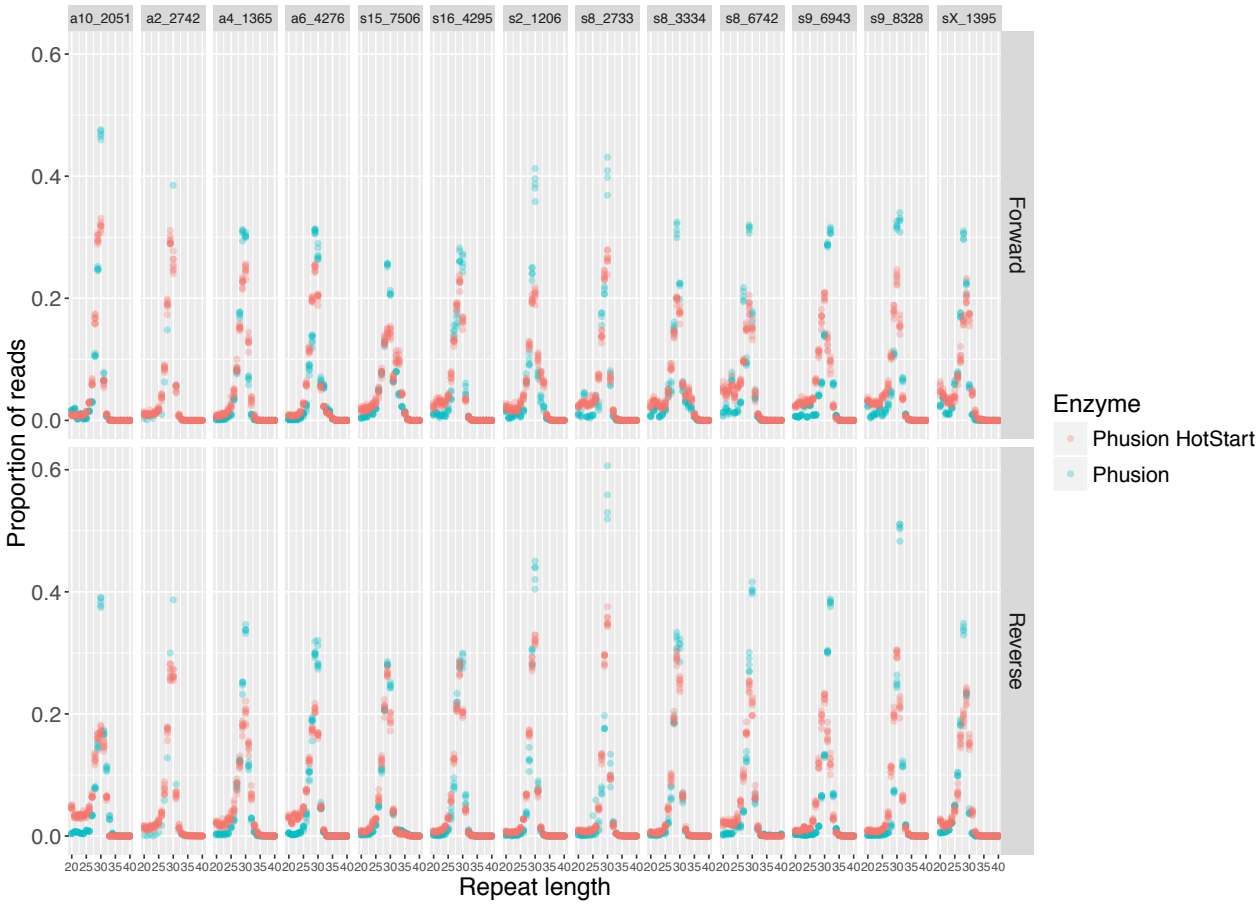

Fig. 3.19 Scatter plot showing the read length distribution following amplification and sequencing in a multiplex PCR of the standard Phusion enzyme compared with the Phusion HotStart. Both enzymes produce highly consistent plots across technical replicates with the standard Phusion producing tighter distributions. However, the Phusion HotStart was favoured as a result of its ability to amplify more primer pairs in multiplex.

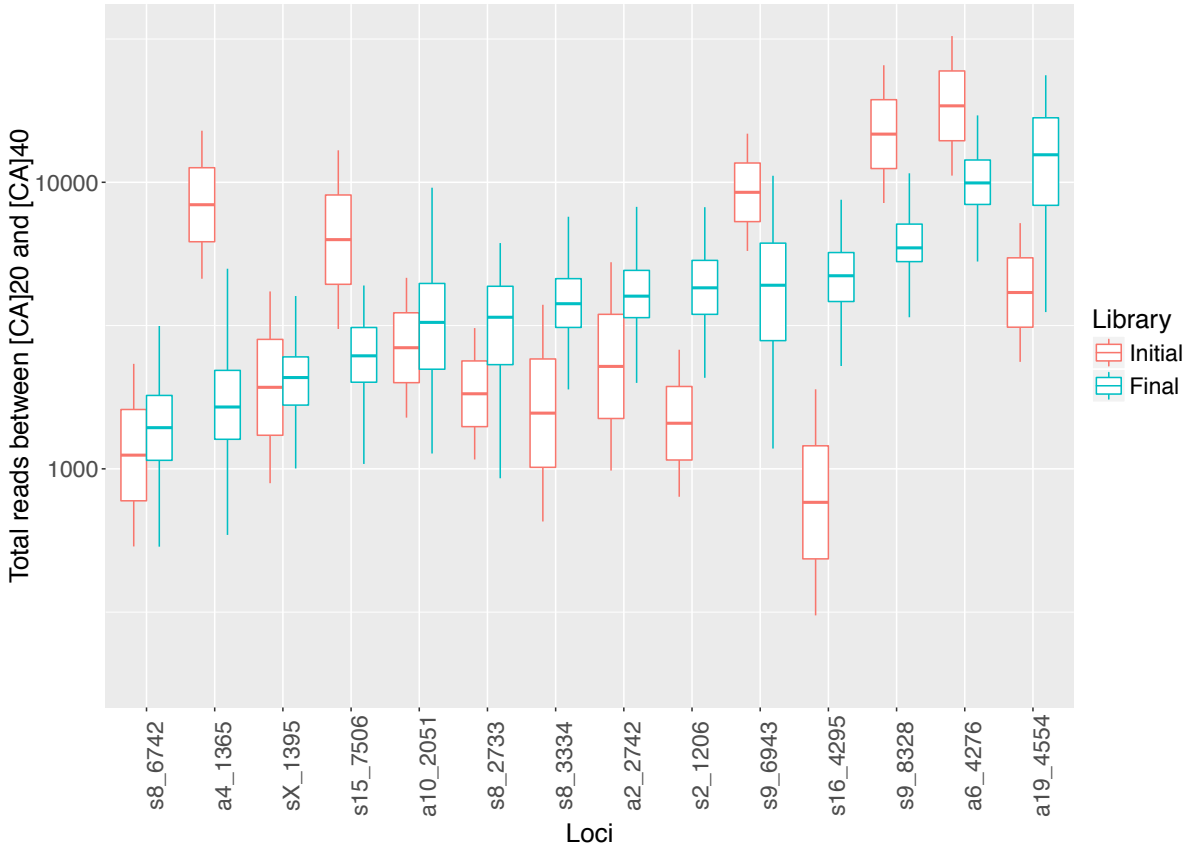

Fig. 3.20 Boxplot displaying spread of read depth across all amplicons in M13_33 before and after primer concentration was optimised to balance amplicon representation.

clone sizes within a single crypt. The protocol has been developed such that the analysis of hundreds of crypts is feasible. Furthermore, based on the data produced by Kozar et al, a multiplex group allowing the amplification and sequencing of over 8 microsatellites should give enough data from approximately 100 crypts to observe age related change in clone size. The final optimised multiplex group (M13_33) amplifies 14 endogenous microsatellites in a single reaction and can be expanded to 15 microsatellite when primer pairs amplifying the transgenic $[CA]_{30}$ microsatellite.

A notable success of this protocol development is the high degree of consistency in polymerase error between technical replicates. This significantly reduces the amount of technical noise generated by the protocol and allows for accurate generation of reference distributions formed from sequencing of reference material e.g. tail clipping or pooled epithelial samples. By generating a reference distribution, any deviation from that distribution observed at a crypt loci would indicate the presence of a mutant microsatellite. Due to the high level of error consistency between reference samples, the protocol should be able to detect relatively small mutant clones, seen in partly populated crypts, that lead to small perturbations from the reference distribution. In addition, the detection of a mutant clone that has drifted towards monoclonality should be easily detectable.

The inclusion of contaminating DNA in the single crypt lysate would lead to under estimation of the mutant clone size due to over represented wild-type DNA. By quantifying the level of cell free DNA present within the fractionation media, it was possible to identify the optimal preparation of the single crypt suspension prior to picking. Furthermore, with the addition of a crypt washing step prior to addition to the lysis buffer, it was possible to visually inspect for any contaminating single cells and ensure that the single crypt alone was transferred. Due to cellular make-up of the crypt, it is near inevitable that there will be a small amount of contaminating DNA and steps can only be taken to account for this contamination during distribution analysis.

The key breakthrough in the development of the sequencing protocol was the discovery of the NEB Phusion DNA polymerase's ability to amplify $[CA]_{30}$ microsatellites with high fidelity and efficiency. Furthermore, the availability of a HotStart version of this enzyme allowed for optimisation of the enzyme for multiplex PCR. A notable observation in the trialling of multiple DNA polymerases was the presence of the Sso7d domain in improving product yield from amplification of low copy genomic DNA. A hypothetical consequence of this is that the increased interaction area between the microsatellite DNA and the DNA

polymerase not only improves processivity but may also improve the enzyme's ability to amplify highly repetitive stretches of DNA. Furthermore, the observed increase in amount of product generated by the Phusion enzyme compared to all the other polymerases tested, suggests a high efficiency in amplification which would theoretically indicate a reduced allele dropout rate. When developing this protocol further in the future, it would be interesting to trial other polymerases that are known to have large interaction areas with target DNA to see if this has an effect on polymerase fidelity at amplifying highly repetitive stretches of DNA.

Each DNA molecule has a certain probability of incurring an error during every round of PCR therefore, reducing the number of PCR cycles is likely to have a substantial effect on the 'blurring' of the read length distribution for a given microsatellite locus. Initial attempts at amplifying $[CA]_{30}$ microsatellites utilised 50 cycles of PCR and showed significant error within the read length distribution. The trial of 25 and 35 cycles of PCR showed a modest improvement in the tightness of the read length distribution when using 25 cycles compared with 35 cycles. This would suggest that later cycles of PCR i.e. cycles 26 to 35 have a minimal effect on the error within the read length distribution with only a small amount of additional blurring seen. When using low copies of template, early errors tend to be over represented within the final amplicon pool which is the likely cause of this reduced effect in later cycles of PCR. The large difference in the read length distribution seen between 35 cycles of PCR and 50 cycles of PCR is likely due to the different polymerases used for the two reactions and is unlikely to be an effect of cycles 36 to 50. In addition to the minimal improvement in read length distribution using 25 cycles of PCR, the use of 25 cycles of PCR was not sufficient to generate a sequencing library alone thus 35 cycles of PCR was chosen as the optimal number of PCR cycles required. When using human crypt material, there is an 8-fold increase in genome copies therefore it would be theoretically possible to use 3 cycles less of PCR (2^3) and generate the same number of amplicons in the final library. Given the relatively minor improvement in read length distribution when reducing the number of PCR cycles by 10, the advantage of a reduction of 3 PCR cycles seems unlikely to be worth pursuing.

Prior to this study, the performance of the Illumina sequencing platform at accurately sequencing microsatellites was unknown. Direct assessment of the sequencing chemistry contribution to the overall error incurred in the microsatellite read length distribution was not performed. However, it is highly likely that there is some contribution as a result of

the Illumina sequencing chemistry. This can be seen when comparing the read length distributions generated between the HiSeq 4000 and MiSeq, that utilise different sequencing chemistry. It would appear that HiSeq sequencing chemistry performs slightly favourably. Furthermore, when comparing the forward and reverse reads at a given locus, the read length distributions differ. This can only be caused by differential sequencing error between the forward and reverse reads as each sequencing cluster represents a clonal representation of a single molecule.

Motivated by the need for consistency between replicates, the majority of optimisation experiments were performed with crypt equivalents generated by diluting reference DNA down to the concentration expected within a single crypt lysate. Once optimisation had been performed, comparison with the read length distributions generated from real crypt lysate compared favourably illustrating the utility of using crypt equivalents as a means of iteratively optimising a single crypt sequencing protocol.

In addition to using crypt equivalents to optimise the protocol, crypt equivalents were used to generate dozens of singleplex reactions targeting different microsatellites in 6 different mice. From this, 5 different loci were shown to be germline variable. It is possible that these loci have a higher mutation rate within the germline and, therefore, may have a higher mutation rate somatically. Though it cannot be ruled out that the variance may be due to contamination from crossing with another strain. Regardless, these 5 loci in the M13_33 multiplex group were included in future analyses to see if these loci do display an increased event rate somatically. As well as being a potential method for identifying somatically mutable loci, this experiment also displays the ability of this protocol to differentiate between highly inbred individuals. This bodes well for the application of this protocol to genetic identification of humans for forensic purposes or to calculate phylogenetic relatedness in population and evolutionary genetics.

Using plasmid cloning of murine genomic loci containing microsatellites, loci containing varying lengths of microsatellite were generated. Using these templates it was possible to show that there is no detectable amplification bias towards microsatellites of shorter lengths. This provides an important control ruling out any possibility of over estimation of clone size when mutant microsatellites are at a shorter length compared to the wild-type length. One caveat to this approach was the 10^6 fold increase in the amount of template copies input into the qPCR reaction when compared with the genomic copy content of a single murine crypt. Thus the possibility that there are biases towards shorter microsatel-

lites in early rounds of PCR that would be exaggerated in low template copy reactions cannot be ruled out.

Using a rational stepwise protocol, the Phusion DNA polymerase was optimised for use in multiplexed amplification of up to 15 amplicons from a single murine crypt. This approach allowed the gradual expansion of the number of primer pairs per multiplex group whilst continuing the use of the high fidelity Phusion enzyme. To try and improve the amplicon balance within each reaction, every time the multiplex group was used in a sequencing experiment, the amplicon balance was assessed and the concentration of each primer pair added to the reaction was adjusted to improve amplicon balance. This dramatically improved amplicon balance and, as a result, reduced the sequencing depth requirements for each crypt.

Intriguingly, the read length distribution for different templates does differ at some loci, for example using microsatellites contained within linearised plasmids compared within genomic DNA generates a far tighter read length distribution from plasmid DNA at locus s9_8328. The source of this difference is currently unknown but could be due to chromatin accessibility at that particular locus in the genomic DNA.

A single murine crypt contains approximately 500 genome copies, the adaption of this technique to a human colonic crypt, which contains approximately 4000 genome copies, should be theoretically possible. The main challenge to overcome in the adaption of this protocol to human crypts is the design of primer pairs amenable to multiplex PCR using the developed protocol. The developed computational pipelines for the identification and design of primer pairs for $[CA]_{30}$ microsatellites are applicable to any reference genome thus should allow for straightforward design of human specific primers.

The motivation for developing this protocol was the application of microsatellite mutations as neutral marks for studying intestinal stem cell dynamics. However, this protocol has potential for wider application. STR genotyping is widely used in evolutionary and population genetics to track relatedness of individuals within and across species. The ability to multiplex different microsatellites from such low amounts of DNA would allow for easy application of this technology to these fields. STR genotyping is also widely used in forensic sciences, particularly using highly degraded samples that are only available at low amounts. The ability of this protocol to work at low template concentrations in a highly quantitative manner with the potential to deconvolute the presence of different microsatellite species at very similar lengths makes this protocol highly amenable to forensic purposes.

Furthermore, this protocol lends itself well to further expansion. The crypt picking protocol currently only allows the isolation of 100-200 crypts per day per person picking. With the use of a microfluidic device or a large particle flow sorting protocol, it may be possible to isolate far more single crypts from the single crypt suspension. It may also be worthwhile exploring whole genome amplification effects on allele bias so as to increase the amount of template DNA to work with. This would allow the use previously published multiplex PCR protocols allowing the simultaneous amplification of 1000s of loci. With such a large array of microsatellites, it would be theoretically possible to infer clone size at a single cell resolution. As some effect of PCR bias cannot be ruled out, inclusion of molecular barcoding to account for any PCR bias would also be a valuable addition to this protocol. Nonetheless, the current protocol provides a powerful method for the accurate quantification of microsatellite length at up to 15 loci in a single murine crypt.

It is next necessary to develop an analysis tool for the detection of mutant microsatellite species. Furthermore, the synthetic loci produced to assess amplification bias in this chapter are ideally suited to mixture experiments. Mixing experiments would allow for *in vitro* validation of the microsatellite sequencing protocol and analysis to detect mutant microsatellites. These validation steps are essential to show that this method has the ability to accurately quantify clone size.

Chapter 4

Development and validation of an analysis pipeline for quantifying crypt clone size from microsatellite sequencing data

As described in the previous chapter, the optimised microsatellite sequencing protocol produces highly consistent read length distributions between technical replicates. This allows for the generation of reference distributions amenable to comparison with crypt loci distributions such that deviations from reference can be interpreted as biologically meaningful. Due to *in vitro* polymerase slippage and sequencing error, the signal generated from microsatellite sequencing is a distribution peaking at the native microsatellite length. The analysis of the data, therefore, must take this distribution into account to allow an accurate inference of the proportion of reads contributing to the distribution that are wild-type and those contributed from mutant microsatellites.

Previously published methods for the analysis of microsatellite sequencing data, utilise either low read depth whole genome sequencing data [37, 45, 49] or use *in silico* designed reference genomes that utilise existing alignment tools to call microsatellite length from targeted re-sequencing experiments [10, 18]. Though the use of tools designed for the analysis of whole genome sequencing data would be able to confidently call the length of the dominant microsatellite, most likely the wild-type $[CA]_{30}$, it would not be useful for detecting minority microsallite lengths. The use of tools for microsatellite length calling from

whole genome sequencing data is better suited to retrospective population analyses that require information from many loci to infer kinship. For the inference of mutant clone size, the contribution of microsatellites of differing lengths to the overall distribution must be inferred. Thus using tools that extract the dominant microsatellite length are not suitable for our needs and a novel tool that allows inference of mutant microsatellite length contribution is required.

The first challenge to overcome in the generation of this tool is to take the raw Illumina sequencing data and generate a matrix representing each crypt, each locus amplified from that crypt and the information generated from both read directions. For each of those data points, each read must be assessed for the length of microsatellite represented. This analysis should lead to a counts matrix that displays the number of reads represented at each possible length of microsatellite, split into a crypt specific, locus specific and read direction specific matrix. The computational method used to generate such a matrix is described in this chapter.

Once a counts matrix is generated, the read distribution resulting from amplification and sequencing error will be observable in the spread of reads seen at different microsatellite lengths. Here lies the second challenge in analysing such datasets. In order to quantify clone size, a computational method for the inference of mutant microsatellite contribution to the overall read distribution is required. The use of mixture modelling and least squares estimation to infer the relative proportions of wild-type and mutant distributions at each locus was opted for. In order to make such estimations, a wild-type distribution and various mutant distributions were required. Wild-type distributions were formed from the sequencing of reference material (such as tail or ear clippings from mice or, in human studies, pooling of colonic epithelium) and mutant distributions were predicted by shifting the wild-type distribution to differing extents to estimate distributions resulting from varying lengths of mutant microsatellite.

To validate this computational approach, as well as estimate the level of sensitivity of the sequencing protocol, a mixing experiment will be performed using the synthetic loci previously described. Using this approach it will be possible to deconvolute the relative proportions of mutant and wild-type microsatellites matched to *a priori* knowledge of the actual mixed proportions. This mixing experiment will provide validation of the sequencing protocol and computational approach for the quantification of mutant clone size. Furthermore, by varying the copies of synthetic loci added to the initial reaction, it will be possible

to assess the effect of copy number on the sensitivity of the method.

4.1 Determining $[CA]_n$ count distribution from FASTQ file

As existing tools for the assessment of microsatellite sequencing data do not quantify relative proportions of microsatellites of differing lengths, a custom tool for counting of [CA] length in each read was required so that length distributions could be ascertained. Furthermore, as loci specific and read direction specific distributions had been observed, it was necessary to have this count data represented as a matrix displaying locus specific and read direction specific information. After sequencing, each crypt sample produced two FASTQ files, one containing forward reads and the other containing the paired reverse reads. Each FASTQ file contained the sequence representation of up to 15 amplicons, depending on the multiplex group used. Each FASTQ file was filtered to ensure 80% of every read was at Q20 or above. A custom Perl script was written to analyse each quality filtered file independently. To separate all of the reads into locus specific groups, the locus specific primer sequence present at the 5' end of each read was used to separate reads into locus specific groups. For each read within a locus specific group, the [CA] repeat length is determined by comparison of the [CA] repeat with a [CA] standard of differing lengths. Once a given read matches the length of the [CA] repeat standard, the [CA] repeat length can be called and the next read is assessed. Through this iteration a read distribution can be generated that is locus specific. As the forward and reverse reads are contained within separate FASTQ files, it is also possible to generate locus specific and read direction specific read distributions, Figure 4.1. Finally, this tool outputs a count matrix that can be used for mixture modelling and least squares inference of relative proportions of wild-type and mutant reads, described in the next section.

4.2 Mixture modelling to infer proportions of $[CA]_n$ mutant/wild-type mixtures

Once a counts matrix detailing locus specific and read direction specific microsatellite length distributions had been generated, a method for determining the proportion of wild-type versus mutant reads was required. One of the most effective ways of achieving this is through predicting mutant clone distributions using mixtures of wild-type data with theoretically

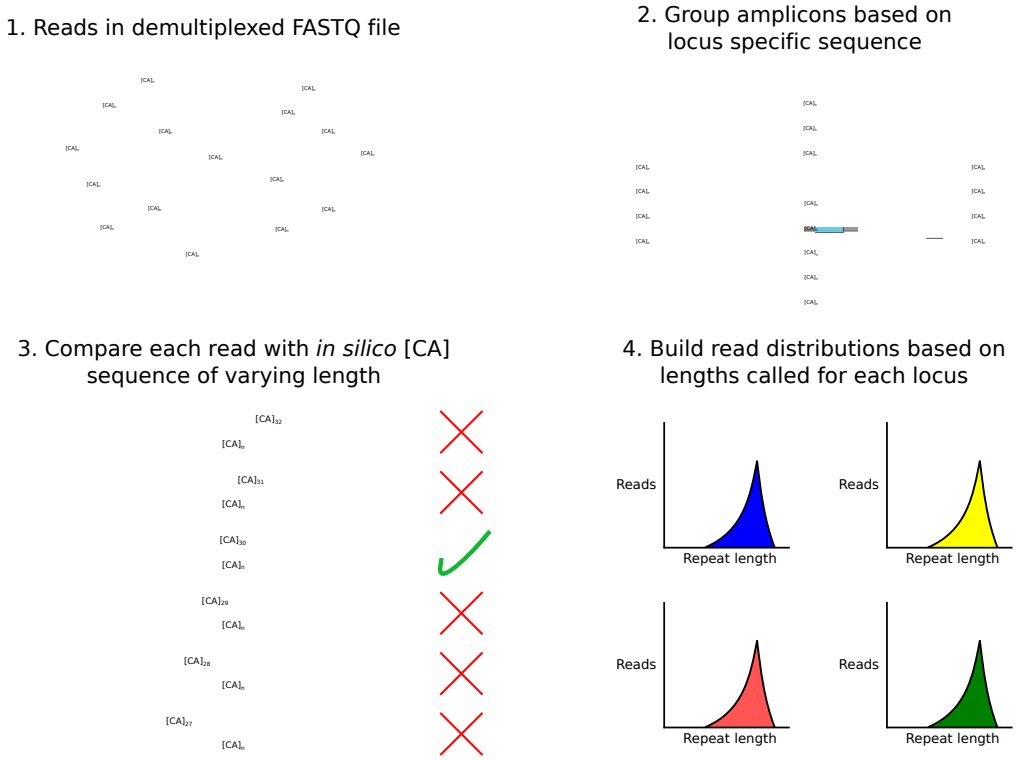

Fig. 4.1 Schematic of process used by the custom Perl script to count the length of [CA] repeats in demultiplexed FASTQ sequencing files. Reads are first grouped based on their locus specific sequences. Each read is then compared with a [CA] repeat standard of varying length to determine the [CA] length present in each read. From this data, read distributions can be generated for each locus present. As the demultiplexing pipeline separates forward and reverse reads, the script generates locus and read direction specific microsatellite length distributions.

predicted mutant data, before comparison with the actual crypt distribution. This mixture modelling approach can be described as such:

$$mix(x, \Phi, shift) = (1 - \Phi)ref(x) + (\Phi)ref(x - shift) \tag{4.1}$$

Therefore, to estimate *mix(x)*, a reference distribution, *ref(x)*, and mutant distribution, *ref(x - shift)*, must be generated. The reference distribution was generated by sequencing reference DNA samples, isolated from ear or tail samples in mice or colonic epithelium samples from humans, diluted to the equivalent of a single crypt. A minimum of 8 technical replicates per reference dataset were performed and the median value from these replicates was used as a reference for downstream analysis. To account for lane to lane variation in reference distribution, reference samples matched to each mouse were included in every lane of sequencing and crypts were only analysed using reference datasets generated on the same lane of sequencing.

This median reference distribution was then shifted to generate mutant distributions representing mutant microsatellites of different lengths. Mixing of a proportion of the mutant distribution (denoted as Φ) with a remaining proportion of wild-type distribution (1 - Φ) simulates the distribution formed from a crypt with a Φ proportion of mutant reads shifted by the value represented as *shift*. This principle is depicted in Figure 4.2.

The predicted model of loop insertion-deletion mutagenesis at these loci leads only to small changes in microsatellite length. Therefore, maximum shifts used were a loss of 5 repeat units or a gain of 5 repeat units. In addition, the use of only small alterations in microsatellite length meant that a shifted wild-type distribution was more likely to accurately represent a mutant read distribution. For example, the polymerase fidelity at a $[CA]_5$ microsatellite is almost certainly going to much higher than that at a $[CA]_{30}$ thus simply shifting the $[CA]_{30}$ distribution 25 units leftward will not accurately simulate a $[CA]_5$ distribution. Instead limiting the shifts to between -5 and +5 shift from the reference distribution, the mixture modelling is likely to be more accurate.

In order to ascertain the shift value and Φ value that most accurately described the distribution generated from crypt sequencing, least squares was used to determine the mixture that most accurately predicts the real crypt data. A least squares value is ascertained by calculating the difference in y-value across a range of x-values, squaring these differences and summing the total. A low least squares value indicates a high similarity between the predicted and actual read distributions whilst a high least squares value indicates the con-

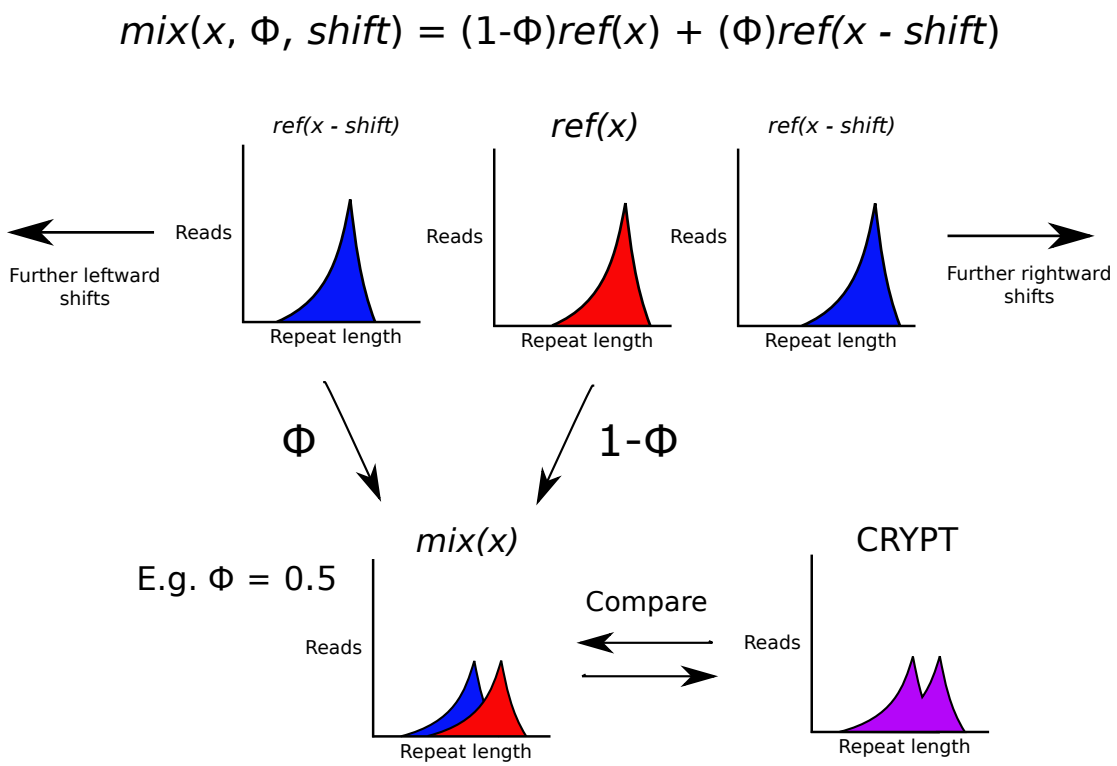

Fig. 4.2 A reference distribution is generated by taking the median value of 8 technical replicates of reference material from the same individual as the crypts were isolated from. Mutated distributions are then predicted by shifting the reference distribution either leftward or rightward. Mutated distributions are then mixed with wild-type distributions at proportions determined by the value Φ. This iterative process generates many different predicted mutant distributions; the distribution that most closely matches the crypt distribution is called as the correct shift value and the correct value of Φ. This information is used to determine if the crypt is partly populated, wholly populated or wild-type. In the example given above, the crypt would be called as having a Φ value of 0.5 with a loss of one [CA] unit.

verse. The shift and Φ value that leads to the lowest least squares value can be inferred as the parameters that best describe the microsatellite length distribution.

4.3 Use of synthetic loci to validate optimised PCR and analysis method *in vitro*

To independently validate the sequencing and analysis method, a mixing experiment was performed using the synthetic loci previously generated in Section 3.10 and summarised in Table 3.5. The mixing experiment simulated samples containing 100% wild-type reads, 75% wild-type reads, 50% wild-type reads, 25% wild-type reads, 10% wild-type reads and 0% wild-type reads, as summarised in Figure 4.3. The template containing a $[CA]_{30}$ microsatellite was always used as the wild-type template with the mutant template varying in length. Each sample was amplified and sequenced with 6 technical replicates.

To test the effect of template copy number on mutant clone size quantification, the mixing experiment was performed with differing amounts of template input. The total number of template copies per sample tested were: 165 copies (equivalent to 1/3 of murine crypt), 500 copies (equivalent to one murine crypt), 4000 copies (equivalent to one human crypt) and 10^6 copies (to test the effect of excess copy number). The results of these experiments are discussed below.

4.3.1 Mixing of reference and mutant microsatellite species to determine limits of mutation detection

A comparison of the Φ estimate with the actual mixed proportions of mutant and wild-type template are shown in Figures 4.4 and 4.5 for mouse crypt equivalent and human crypt equivalent respectively. Even at only 500 plasmid copies, there is a good agreement of the Φ value and the actual mixed proportions at both loci, at 5 differing lengths of mutant loci and at a wide range of mixed proportions. This clearly shows the ability to detect clones of varying sizes, regardless of mutant microsatellite length, at template copies equivalent to a single murine or human crypt.

Locus a4_1365 shows particularly good agreement between Φ estimate and known mixed proportions. On the other hand, locus s9_8328 agrees less well, particularly at mixtures with lower levels of wild-type DNA. This error is a result of the assumption that a shift

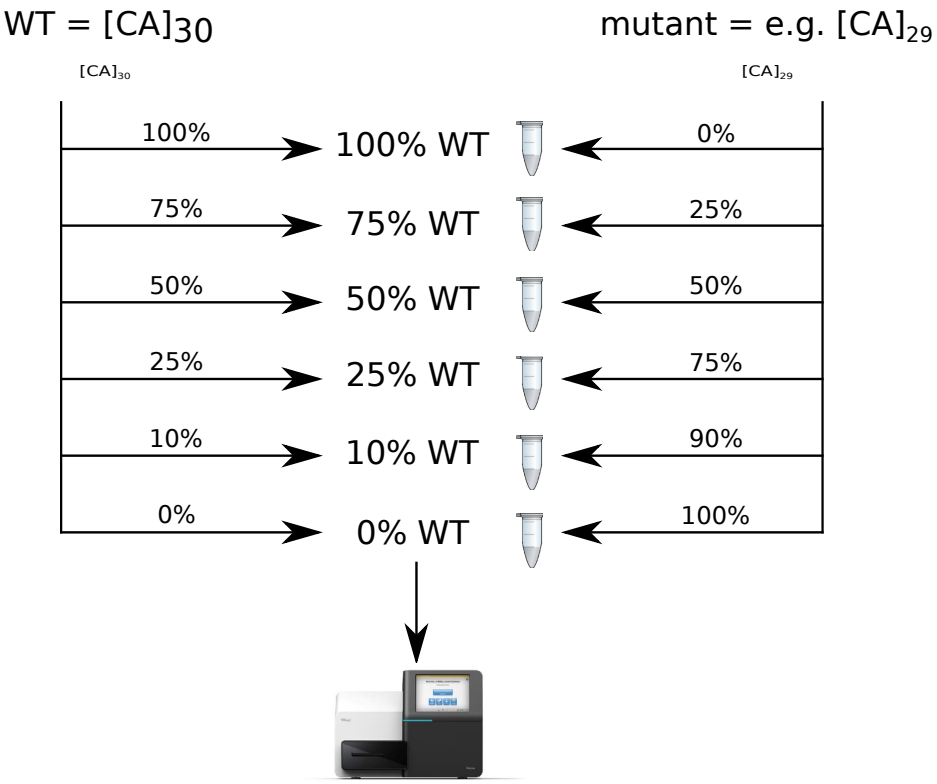

Fig. 4.3 Schematic of the protocol used to simulate partly mutated crypts with synthetic loci. This was done with 7 different lengths of microsatellite representing two different genomic loci. $[CA]_{30}$ was always used as the wild-type template with the length of the mutant template being varied.

in the reference distribution would accurately predict a mutant microsatellite distribution. This can be seen when the length distributions for the pure plasmid populations are plotted, Figure 4.6 . It can be seen that there is a slight change in the shape of the read distribution as the microsatellite length changes. This is most likely due to a change in the polymerase and sequencing error rate as a result of a change in microsatellite length. The change in distribution shape with length is particularly pronounced at locus s9_8328, where the largest error in Φ estimates is observed. In order to accurately interpret each Φ value, it is going to be necessary to take these observations into account. Methods to deal with this disparity are discussed in Section 4.5.

4.3.2 Variable copy number in synthetic loci mixtures reveals stochastic effects as a source of estimation variation at low copy number

To ascertain the effect of template copy on accurate clone size quantification, the mixing experiment described in Section 4.3.1 was repeated with lower template copy (165 copies) and higher template copies (10^6 copies). To assess the level of variation seen at different plasmid copies, the mean standard deviation in Φ estimate was calculated across all mutant lengths and all technical replicates in the mix containing 50% wild-type and 50% mutant template. As can be seen in Figure 4.7, when approximately 500 plasmid copies are added to a reaction, the mean standard deviation is very low (0.040 at a4_1365 and 0.033 at s9_8328). However, when 165 plasmid copies are added, the mean standard deviation is far higher (0.140 at a4_1365 and 0.143 at s9_8328) indicating a large increase in Φ estimate variability at plasmid copy numbers close to 165. Massive excess of template does not have the same effect thus the most likely cause of variability at low template copies is allele dropout. It can be concluded from this analysis that the template copy input at which allele dropout begins to have a large effect is between 165 and 500 plasmid copies with large amounts of DNA having little effect on estimate variation.

4.3.3 Stochastic error in 50:50 mixes of low copy wild-type and mutant templates supports absence of amplification bias

Previous analysis in Section 3.10.2 showed no bias in PCR amplification of microsatellites of different lengths. However, due to the addition of 10^6-fold more plasmid copies to the qPCR reaction than what would be seen in a single murine crypt reaction, the possibility of

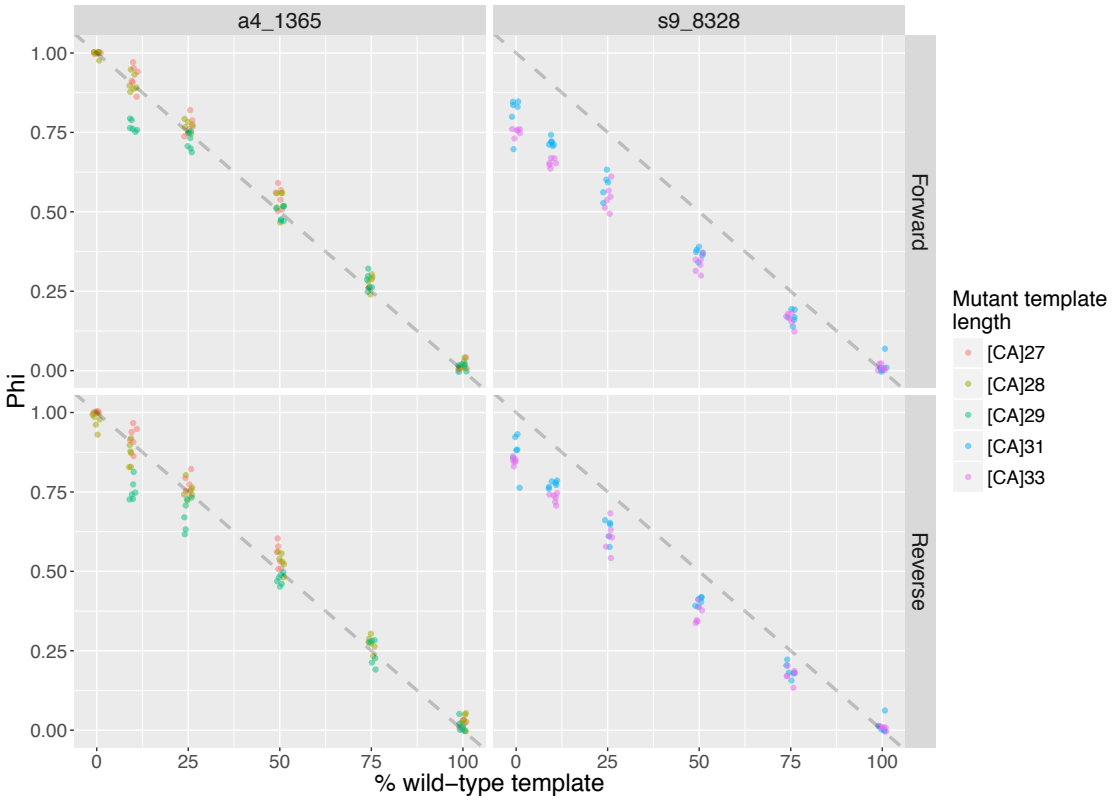

Fig. 4.4 Jitter scatter plot showing the estimated Φ values from 6 technical replicates of plasmid mixing at 0% wild-type, 10% wild-type, 25% wild-type, 50% wild-type, 75% wild-type and 100% wild-type template proportion. In each reaction, there was 500 plasmid copies, equivalent to the number of genome copies expected in one murine crypt.

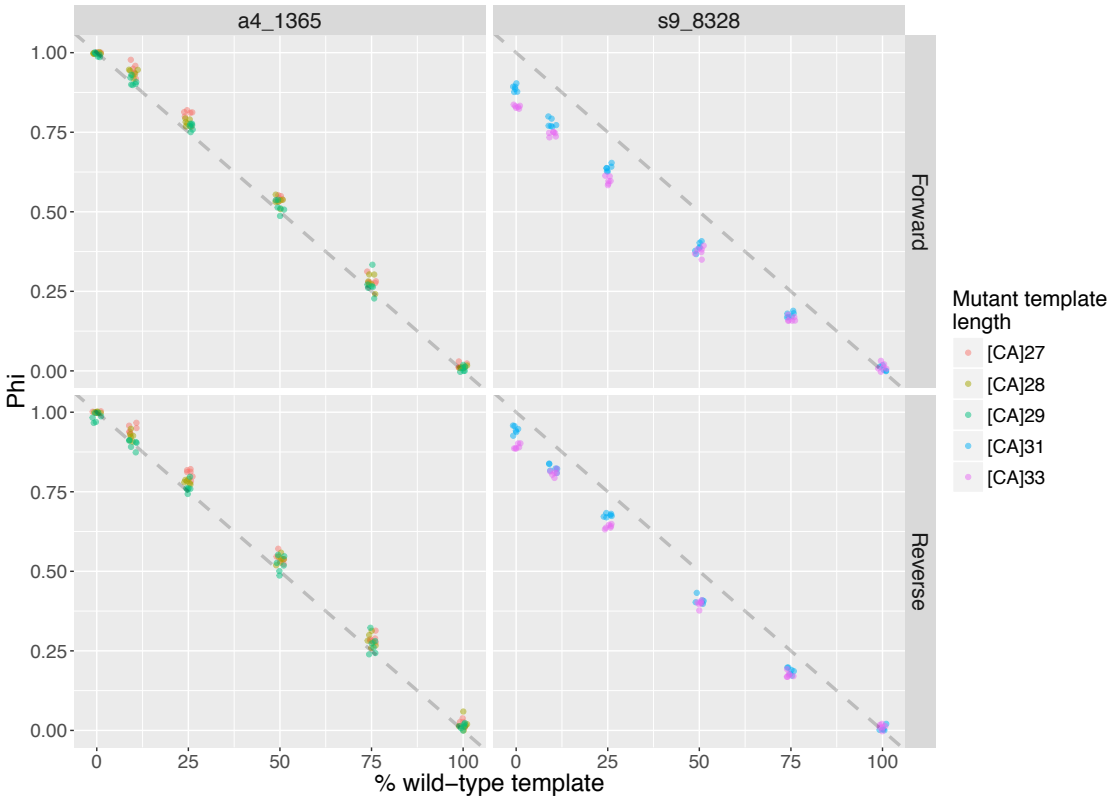

Fig. 4.5 Jitter scatter plot showing the estimated Φ values from 6 technical replicates of plasmid mixing at 0% wild-type, 10% wild-type, 25% wild-type, 50% wild-type, 75% wild-type and 100% wild-type template proportion. In each reaction, there was 4000 plasmid copies, equivalent to the number of genome copies expected in one human colonic crypt.

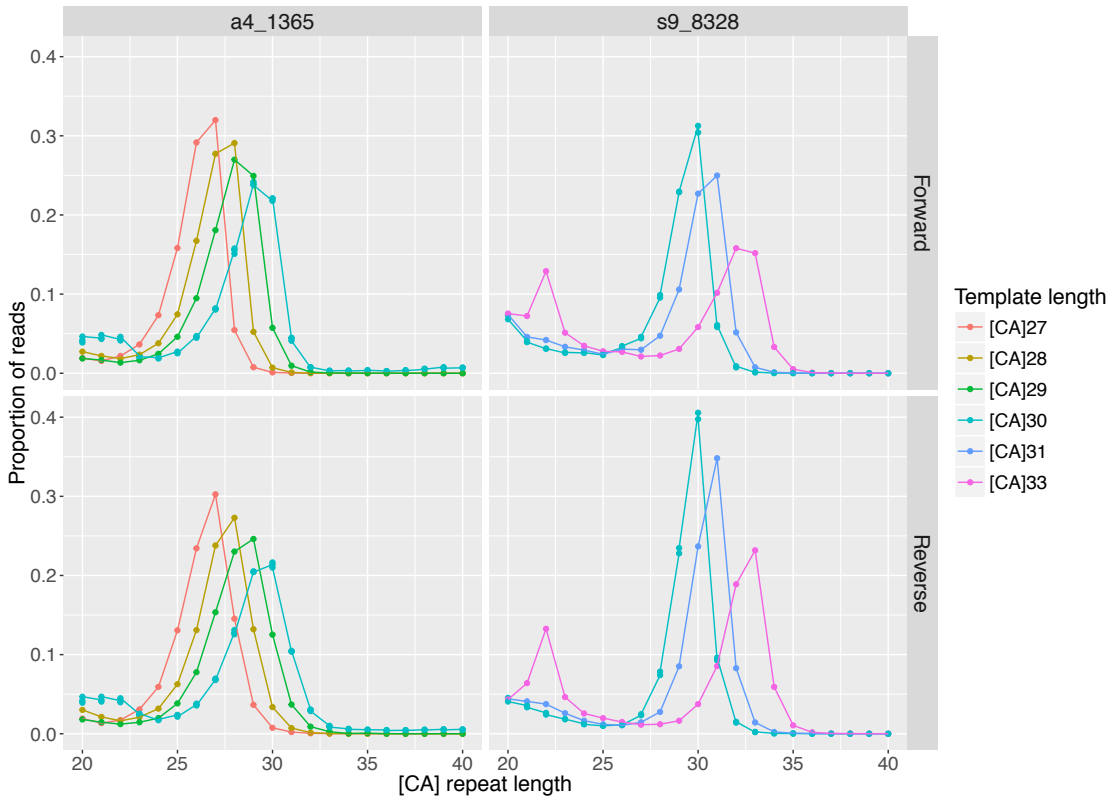

Fig. 4.6 Scatter plot with line annotation showing mean read proportion of read distributions of 7 different lengths of $[CA]_n$ microsatellites representing two different genomic loci. It can be seen that at the same locus in the same read direction, the read distribution differs. This is likely to be due to changes in polymerase and sequencing error rate for different lengths of microsatellite. As a result using a shifted reference distribution as simulation of a mutant microsatellite distribution leads to a slight inaccuracy in Φ estimates.

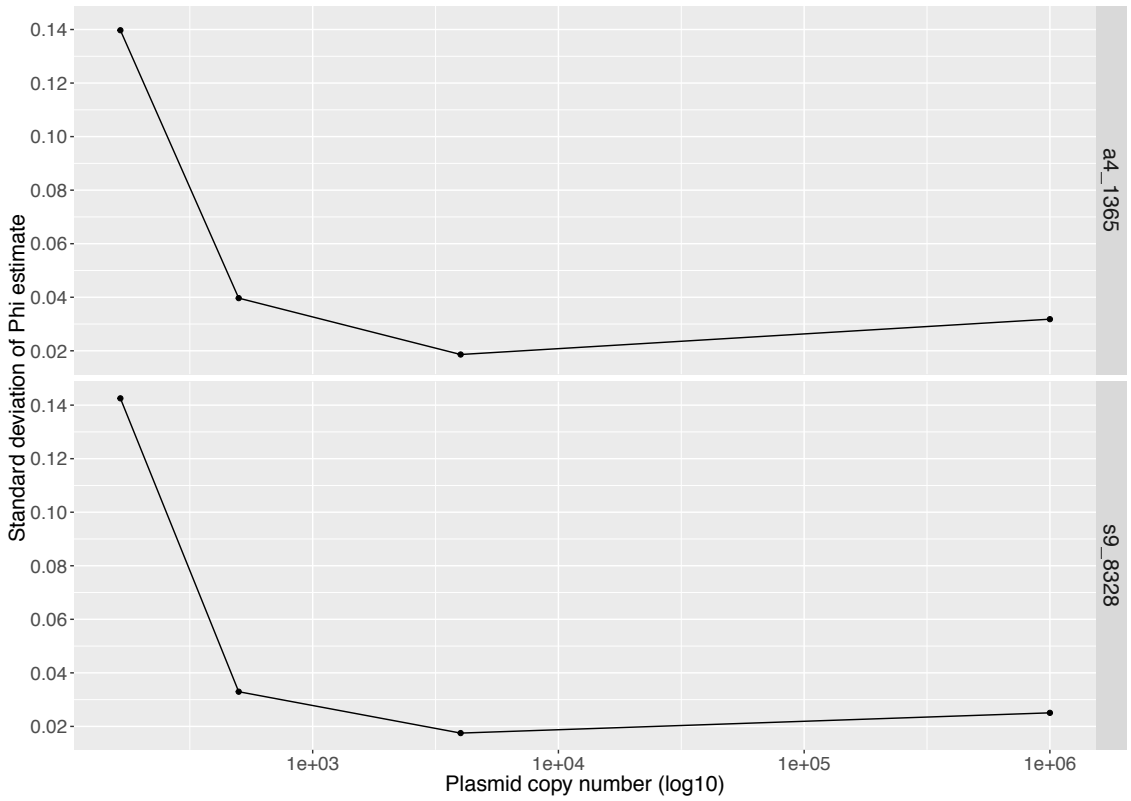

Fig. 4.7 Scatter plot with line annotation showing standard deviation of Φ estimate with changing amount of template input. Each standard deviation was calculated from 50% wild-type template mixtures and averaged across all technical replicates and all different mutant template lengths. As the number of template copies is reduced below 500 copies per reaction, the standard deviation of the Φ estimate significantly increases. It is likely that this is due to allele dropout effects.

amplification bias at low template copies could not be ruled out. An interesting observation from experiments containing 165 plasmid copies was the consistent nature of the variability around the median, Figure 4.8. It is highly suggestive that the allele dropout effects are entirely stochastic and there is no bias towards microsatellites of different lengths, at least between microsatellite length changes of -3 repeats and +3 repeats which are represented in this experiment. This would again indicate that allele dropout should have negligible effects on clone size estimates and the largest sources of error are from read distribution changes when microsatellite length shifts from the wild-type length. Furthermore, if a crypt lysate were split into two halves and amplified separately, though the number of plasmid copies entering the reaction would be reduced, two Φ estimates would be ascertained the average of which is likely to be an accurate estimate of Φ.

4.4 Validation of crypt washing as an appropriate method for minimising DNA contamination

The plasmid mixing experiments demonstrated the ability to deconvolute proportional mutant microsatellite *in vitro* using synthetic loci. To demonstrate the ability at detecting mutant clones in an *in vivo* setting, the Rosa26-$[CA]_{30}$-eYFP mouse was used to identify crypts containing a mutation in the transgenic $[CA]_{30}$ microsatellite by observation of YFP positivity. Using a dissecting microscope with a fluorescent bulb, it was possible to identify YFP+ crypts during micropipette isolation, Figure 4.9. Primers were designed to amplify the transgenic $[CA]_{30}$ microsatellite within the Rosa26-$[CA]_{30}$-eYFP construct in a singleplex reaction. In Figure 4.10, it can be seen that there is a clear distinction between YFP+ and YFP- crypts. However, the variability of Φ estimates for the YFP+ crypts was unexpected. As this variability was not observed in the YFP- crypts, it was hypothesised that the variability may be due to wild-type DNA contamination in the YFP+ crypt lysate.

To test whether wild-type DNA contamination was leading to YFP+ Φ variability, the experiment was repeated but each single crypt was washed 4 times in PBS before transferring to the lysis buffer (Section 2.8.5). As can be seen in Figure 4.11, this had a significant effect on the level of variability in Φ estimates for the YFP+ crypts. As a result, all further crypt picking utilised this washing technique to minimise wild-type DNA contamination.

There were two unexpected results from this experiment. Firstly, the YFP+ WPCs do not lie at a Φ value of 1 and cluster around a Φ value of approximately 0.8. Though this

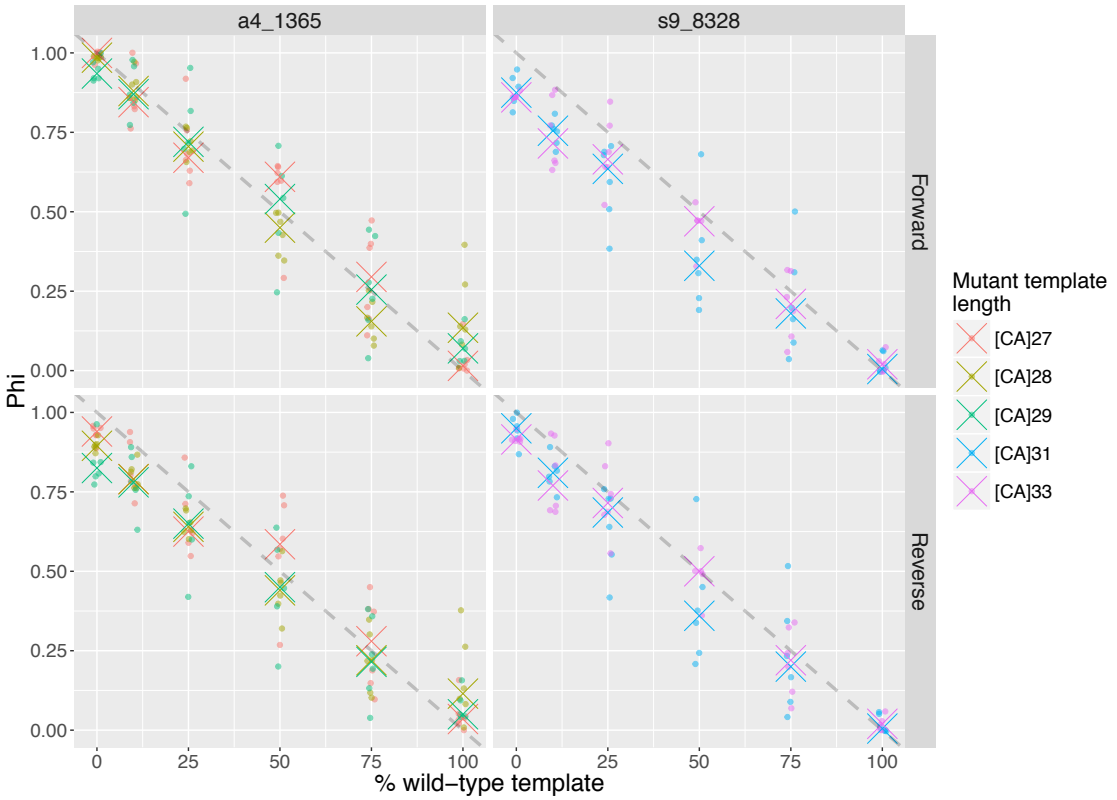

Fig. 4.8 Jitter plot showing technical replicates of low template copy reactions (approx. 165 copies per reaction) at varying proportions of mutant to wild-type and varying mutant microsatellite length. The median Φ estimate is represented by the cross. Individual replicates are shown varying either side of the median with no distinct pattern. This random variation in error is suggestive of a stochastic process determining whether either the wild-type or mutant microsatellite predominates. This supports the absence of amplification bias and suggests that taking multiple estimates of Φ using more than one low template reaction could give an accurate estimate of overall Φ despite individual reaction error.

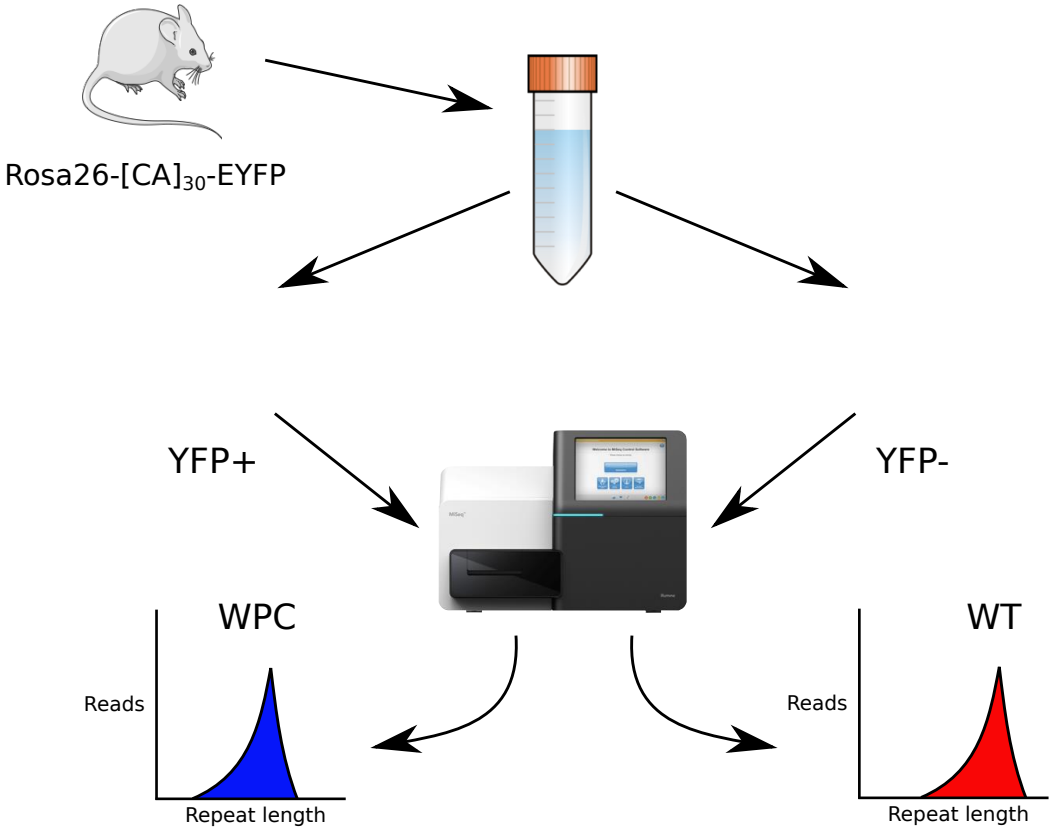

Fig. 4.9 Schematic showing method used for picking YFP+ crypts for comparison of read distribution with YFP- crypts.

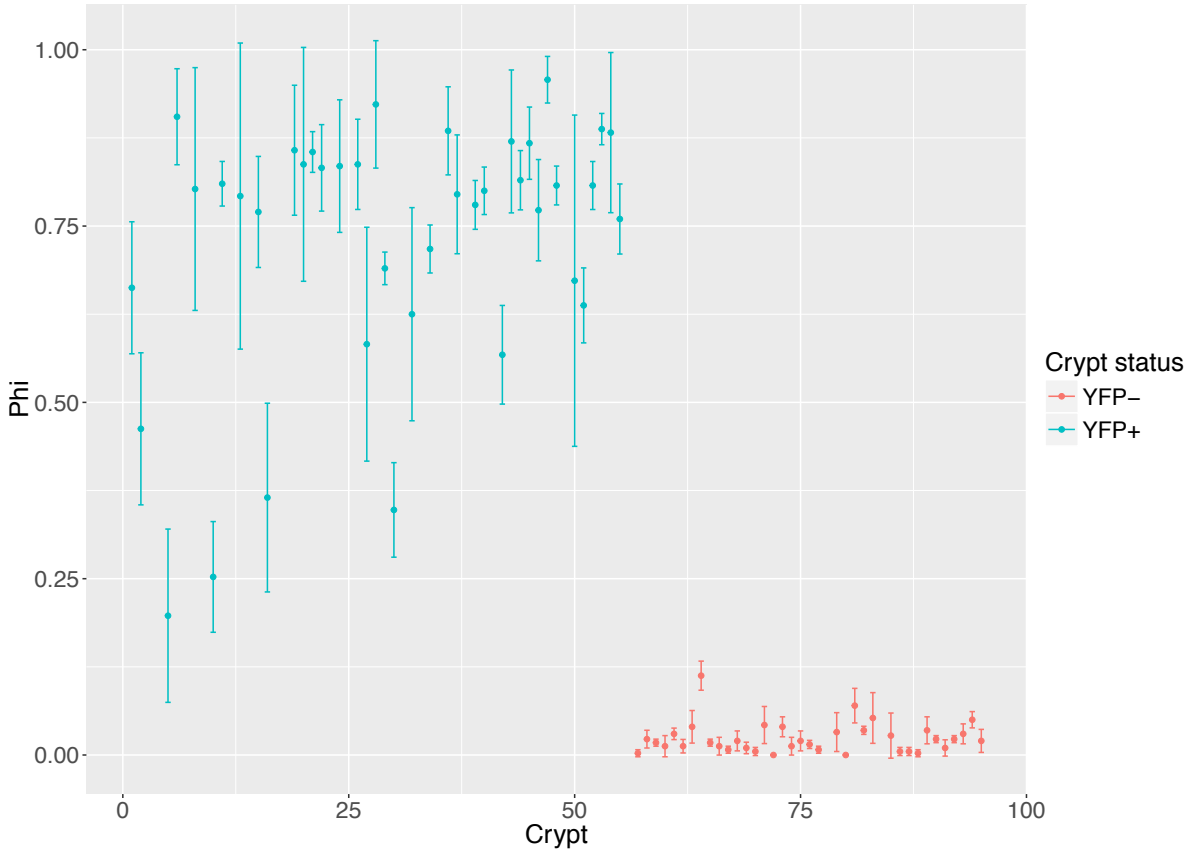

Fig. 4.10 Scatter plot with error bars indicating standard deviation of Φ estimates of YFP+ and YFP- crypts. Without PBS washing each crypt, the error is very large for YFP+ crypts and estimates are far from a Φ value of 1, as would be expected from YFP+ crypts.

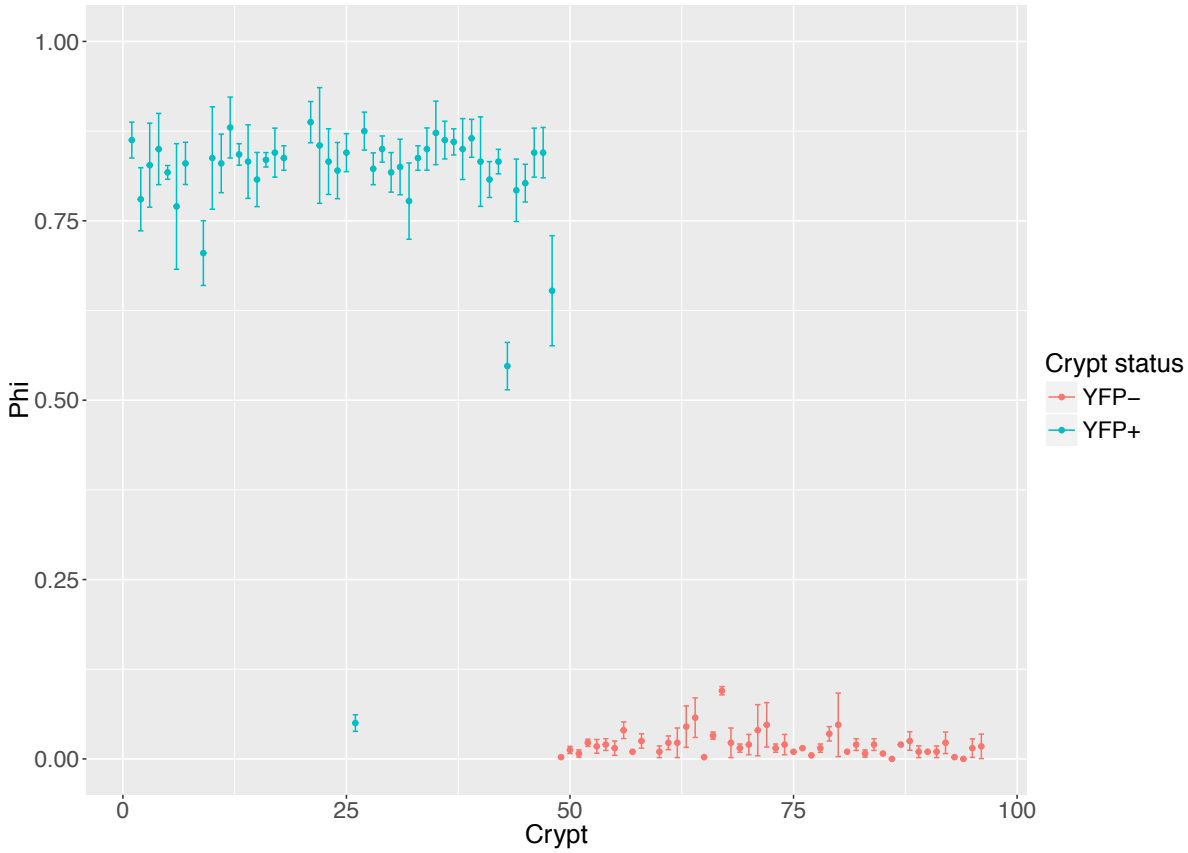

Fig. 4.11 Scatter plot with error bars indicating standard deviation of Φ estimates of YFP+ and YFP- crypts. With PBS washing of each crypt, the level of error between crypts is much lower when compared with Figure 4.10 and the Φ values are much closer to 1 in YFP+ crypts, as would be expected.

clearly distinguishes them from the wild-type crypts that remain close to a Φ value of 0, in an experiment with no *a priori* knowledge of clonal status, it would be difficult to differentiate between WPCs and PPCs, particularly those containing large clones. There are likely to be two main sources of this error: 1) there is an inevitable amount of wild-type DNA that will contaminate any mutant crypt as a result of transfer of stromal cells. 2) The mixture model relies upon an assumption that the predicted mutant distribution takes the exact shape of a shifted reference distribution. As can be seen in Figure 4.12, though the distributions are relatively similar in shape, there is a difference that would mean that a shifted reference distribution would not exactly predict the mutant distribution but a contribution of wild-type stromal cells cannot be ruled out. Overall, it would seem that the error in Φ estimate is a result of distribution shape change at different lengths of microsatellite and an inevitable amount of wild-type DNA contamination from associated wild-type cells. This inaccuracy can be accounted for by estimating the Φ value generated by WPCs at a given locus, for example at the YFP locus a Φ value of 0.75 and above is likely to be a WPC, and a value of less than 0.1 is likely to be a wild type crypt with any crypt with a Φ value between 0.1 and 0.75 likely being a PPC. Thus this first unexpected result can be accounted for whilst interpreting the Φ estimates for each locus.

The second unexpected result was the presence of 3 anomalous Φ estimates (Crypt 26, 43 and 48). It was expected that these crypts have a Φ value of 0.75 or higher, based on their YFP status, instead Φ values consistent with a wild-type crypt (Crypt 26, Φ = 0.05) and two PPCs (Crypt 43 and 48, Φ = 0.55 and 0.65 respectively) were inferred. As can be seen in Figure 4.12, the Φ values inferred appear correct and the anomalous result is likely due to a real signal difference and is not a consequence of a flaw in the computational approach. The YFP+ crypts are picked under a low power dissecting microscope that only observes the crypt in one plane thus it is possible to miss small unmarked YFP- clones in a background of YFP positivity. It is, therefore, likely that the two crypts with a Φ value of between 0.1 and 0.75 were correctly called as PPCs. The YFP+ crypt called as having a Φ value of 0.05 is possibly a YFP- crypt picked in error. Alternatively, a small false negative rate may be present in the protocol. Regardless, the error appears to be present at a low level and should not have a major influence on the overall power of this protocol in quantifying intestinal stem cell dynamics.

An additional interesting observation from sequencing YFP+ crypts was that all mutations were caused by a +1 expansion in microsatellite length. This is consistent with the

predicted loop insertion-deletion mechanism of microsatellite mutagenesis leading to small scale changes in microsatellite length with large scale changes being far rarer. This further supports the decision to only use small scale changes in microsatellite length as clonal marks.

4.5 Discussion

In this chapter, I describe an alternative computational method for the calling of microsatellite length from high read depth targeted re-sequencing data. In contrast to previous methods for calling microsatellite length from sequencing data, this approach allows for the building of a reference distribution that is both locus and read direction specific. Furthermore, by using a mixture modelling and least squares analysis it is possible to detect mutant microsatellites within a mixed population. Previous methods produce summary data that is not conducive to such analysis. To test the combined amplification, sequencing and analysis protocol, various mixing experiments were performed using synthetic loci of known microsatellite length to validate the method as a feasible means of identifying mutant microsatellites within a population composed of two different microsatellite lengths. It was also shown that this method can be used to identify mutant clones *in vivo* and confirm that crypt washing is a necessary step to ascertain accurate clone size estimates.

The development of a computational method that directly utilises the FASTQ file generated by Illumina sequencing for the building of loci and read direction specific read distributions allows for rapid analysis of microsatellite sequencing data. Through the use of loci specific sequences, the counting script is readily scalable to much larger panels of primer pairs. In addition to the feasibility of scaling for [CA] repeats, this analysis method is readily transferable to the analysis of any sequence repeat e.g. STRs commonly used for identification in forensic sciences or STR loci used for study of evolutionary dynamics in ecology. This method of microsatellite length calling is flexible, scalable and has the potential for wide application.

The use of mixture modelling allows for the generation of predicted mutant mixed distributions which can then be used to infer the mutant microsatellite contribution (Φ) and the number of [CA] repeat units gained or lost in the mutant microsatellite (*shift*). This analysis is particularly important given the need to account for distribution generation during microsatellite amplification and sequencing. Though the optimised microsatellite sequenc-

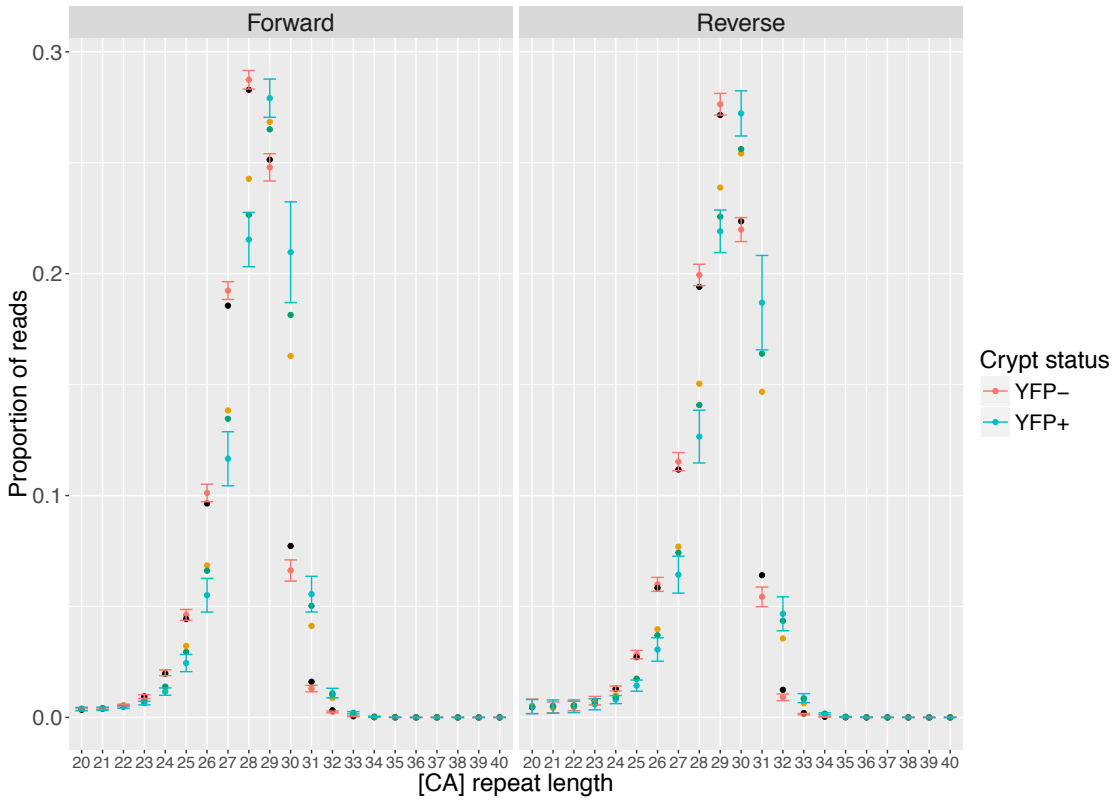

Fig. 4.12 Scatter plot displaying YFP+ and YFP- reference distributions with error bars indicating standard deviation of the estimates. Three further read distributions are indicated in black (Crypt 26, $\Phi = 0.05$), green (Crypt 43, $\Phi = 0.55$) and orange (Crypt 48, $\Phi = 0.65$). The method of calling a Φ value would appear correct and the read distributions represent a real signal change likely cause by human error and/or inability to observe small YFP- clones during picking.

ing protocol has improved fidelity and replicate consistency, the need to account for length distribution is essential for accurate estimate of mutant microsatellite populations within a single crypt sample. The development of a mixture modelling method combined with a least squares inference of mixture parameters, allows for the accounting of such distributions and a significant improvement in mutant clone size estimations. Furthermore, these models are well described, routinely used and can be easily implemented.

Mixture modelling relies upon the assumption that a shifted reference distribution accurately predicts a mutant distribution. As presented in this chapter, this is not always true and the mutant distribution shape can change as a function of the length change, at some loci. Generally, mutations leading to an expansion of the microsatellite length generated distributions representing greater error with a wider, less peaked distribution whilst mutations leading to a contraction in microsatellite length generated tighter, more peaked distributions. Based on previous data suggesting that small loop insertion-deletion events are the main drivers of microsatellite mutation, it would be predicted that the majority of mutations observed in $[CA]_{30}$ microsatellites will be small scale changes. Thus restricting observations of mutations to between $[CA]_{25}$ and $[CA]_{35}$ is unlikely to significantly reduce the number of events observed. Furthermore, by restricting the length change mutations used as clonal marks to between a gain of 5 [CA] repeats and a loss of 5 [CA] repeats, the wild-type and mutant distributions are likely to have similar distributions, further reducing any Φ estimate inaccuracy.

Nonetheless, it is still necessary to account for this slight inaccuracy in Φ estimation when interpreting the data. There are two possible ways to do this: 1) In older individuals, the number of WPCs will be far higher than the number of PPCs. It would, therefore, be possible to isolate distributions with the highest Φ values and, in essence, build mutant distributions. Using these mutant distributions would allow for more accurate to Φ values to be inferred. 2) The thresholds used for differentiating between wild-type, PPC and WPC can be adjusted on a locus by locus basis. For example, the Φ cut off for a wild-type crypt could be 0.1 or lower, the cut off for a WPC (on an X-linked locus in a male) could be 0.9 or higher and anything in between would be categorised as a PPC. These cut offs can be informed by relatively small datasets such that young individuals will have majority wild-type crypts and old individuals will have a substantial amount of WPC allowing for the data to inform threshold values. Using one of these two methods would account for the error generated by length distribution change when shifted. However, it should also be noted that

at locus s9_8328, where significant changes in read distribution were observed, it was still possible to differentiate between clones of different sizes with a high degree of accuracy such that the presence of this error does not substantially affect the ability to identify mutant clones. Thus suggesting that adjustment of Φ thresholds alone will be appropriate for accurate interpretation of clone status.

As further sequencing of these loci is performed, a bank of mutant distributions for multiple loci can be developed such that estimates of clone size can continuously be improved as more data is generated. Once these large *in vivo* datasets are generated, it would be interesting to study these mutant distributions further to see if generic rules can be formulated to accurately describe how a microsatellite read distribution will change as a function of its length thus allowing for accurate generation of mutant distributions for loci with no previous data available. Therefore, adjustment of Φ thresholds is suitable in the short-term and improved Φ estimates with concomitant adjustment in Φ thresholds will improve sensitivity and accuracy in the long-term.

The use of synthetic loci to perform mixing experiments was invaluable in ascertaining the sensitivity of the current protocol. By performing mixing at various proportions of wild-type reads, with as few as 500 plasmid copies, it was possible to distinguish clones differing by as little as 10% wild-type proportions across a large range of plasmid copies (500 to 10^6 copies). Stochastic allele dropout effects are only clearly observed when plasmid copy number is reduced to 165 copies; defining the lower limit of the assay for accurate clone size quantification. Even at this low amount, the difference between mutant and wild-type samples is clear suggesting that for simply ascertaining microsatellite length differences the assay could be used at lower copy numbers. Overall, the use of synthetic loci mixing provided *in vitro* validation of the method at accurately quantifying mutant population size in samples containing copies equivalent to a single murine crypt and a single human crypt.

Previous attempts at revealing any amplification bias towards shorter microsatellite lengths using qPCR was confounded somewhat by the need for larger copy numbers. The use of 50:50 mixes of wild-type and mutant microsatellites at low copy number (165 plasmid copies per reaction) revealed significant variability suggesting that allele dropout is present when the copy number is reduced. An interesting observation in this experiment was the Φ estimates appeared to cluster around 0.5 with no obvious trend towards error above or below 0.5. This would suggest that allele dropout is entirely stochastic and that no amplification bias, within the microsatellite lengths tested, is present at low template copy. This is

consistent with the qPCR analysis performed in Section 3.10.2.

In addition to an *in vitro* validation of the method, the Rosa26-$[CA]_{30}$-eYFP reporter mouse was used to isolate YFP+ WPCs providing *a priori* knowledge of the presence of a microsatellite mutation. 93 out of the 96 crypts picked matched the *a priori* knowledge of mutation status providing *in vivo* validation of the method. The 3 crypts that generated spurious results are likely to result from experimental error though the possibility of a low false negative rate cannot be ruled out. As well as providing *in vivo* validation of the protocol, this experimental approach was used to observe the effect of crypt washing in the improvement of clone size estimates. The Rosa26-$[CA]_{30}$-eYFP reporter mouse is thus a valuable tool in the development of microsatellite sequencing protocols for clone size estimation.

Using YFP+ WPC sequencing a highly uniform mutation spectra was observed. Out of 48 crypts picked in the washing experiment, 47 had an expansion of microsatellite length of one [CA] repeat and the one crypt that did not have an expansion mutation was inferred as having no mutation at all and was likely to be a YFP- crypt picked in error. As all the mutations lead to YFP expression, the microsatellite length change by definition had to generate an in-frame length change thus the range of mutations was somewhat restricted. Even with this in mind, it would appear that the observation of only one mutation event in all YFP+ WPCs suggests a highly stereotyped mutational process most likely described by small loop insertion-deletion events and restricted to stepwise, restricted mutation events with large-scale mutation events being far less likely. This would suggest that the mutation rate estimated by Kozar et al [54], though it will be twice as high at endogenous loci due to biallelic presence compared with monoallelic presence in the transgenic mouse, may not be greatly increased as a result of being able to observe mutations that do not lead to 'in-frame' mutations. This would further suggest that using small scale microsatellite changes is a reasonable method of tracking clones *in vivo*.

Overall, I have provided clear validation of the sequencing and analysis protocol for the accurate quantification of clone size from low template copy samples. Synthetic loci and the Rosa26-$[CA]_{30}$-eYFP reporter mouse were used for *in vitro* and *in vivo* validation respectively. Through setting of accurate thresholds for differentiating between wild-type, PPC and WPC, it will be possible to infer the proportion of PPCs and WPCs in mice at differing ages. The next step in the development of this method is to show that the expected intra-cryptal clone sizes can be observed through use of microsatellite sequencing alone.

Chapter 5

Identifying intra-cryptal clone size variation in mouse colon using microsatellite sequencing

In this chapter, I aim to show that the expected intra-cryptal clone sizes can be observed using microsatellite sequencing. Previous experiments, in this dissertation, have shown that microsatellite sequencing is able to accurately quantify clone size *in vitro* and *in vivo*. As the age related clone size changes in the mouse colon have been well described, it will be possible to validate microsatellite sequencing as a means of detecting these clone size changes in a transgene-free, label-free manner.

As was discussed in Chapter 4, the mixture modelling inference relies upon the assumption that a shifted reference distribution accurately predicts a mutant distribution. Though the model robustly identifies mutant clones, interpretation of Φ values has to be adjusted on a locus by locus basis to appropriately interpret the Φ value inferred by the model. In this chapter, the approach used to define these thresholds is described.

Once set, these thresholds can then be used to interpret microsatellite sequencing of single crypts and allow for the clone size and incidence to be established, in wild-type and genetically altered epithelium. Firstly, clone incidence and size will be established in wild-type mice to demonstrate the feasibility of the method in identifying intra-cryptal clone size differences. Secondly, clone incidence and size will be established in mice with mismatch repair deficiency. The increase in mutation rate associated with mismatch repair deficiency will allow for observation of age related clone size change in a small cohort of mice. Us-

ing simulations, the clone size distributions for different microsatellite mutation rates can be predicted and used to validate observations made in murine colon with mismatch repair deficiency. Furthermore, by comparing the clone size distributions observed using microsatellite sequencing with the simulated data, it will be possible to infer the microsatellite mutation rate in mismatch repair deficient epithelium.

In chapter 3, five loci were identified as being potentially germline variable, Figure 3.11. Sequencing of single crypts will allow for observation of any somatic mutability observed at these loci.

Finally, it has been hypothesised that microsatellites mutate via a loop insertion-deletion mechanism. This mechanism leads to small alterations in microsatellite length with large scale alterations rarely observed. As this method allows for observation of the mutational spectra observed at these loci somatically, it will be possible to see if the mutational spectra observed are consistent with a loop insertion-deletion mechanism of mutagenesis in both wild-type and mistmatch repair deficient epithelium.

5.1 Setting Φ value thresholds for interpretation of intra-cryptal clone size

As discussed in Chapter 4, at some loci, there is a slight change in the shape of the distribution as the length of the microsatellite changes. As the mixture modelling method used to infer Φ relies upon predicting mutant distributions based on a shifted reference distribution, this change in shape leads to inaccuracies in Φ estimation. This effect can be accounted for by adjusting the value of Φ used as a threshold for calling a crypt as wild-type, partly populated or wholly populated. As the extent of these changes is locus specific, it is necessary to adjust the thresholds on a locus by locus basis.

5.1.1 Loci heterozygous for microsatellite length cannot be used for Φ estimation

The mixture modelling method also assumes that, initially, a single length of microsatellite is present at each locus thus any biallelic locus must be germline homozygous for microsatellite length. It is, therefore, necessary to remove any loci that are germline heterozygous for microsatellite length. This was done by visual inspection, on a mouse by

mouse basis, of data obtained from sequencing of reference material e.g. a tail clipping. An example of reference samples prior to filtering is shown in Figure 5.1. In the mouse shown in Figure 5.1, locus s15_7506 was excluded due to disparity between forward and reverse reads, highlighted in Figure 5.2, the rest of the loci were homozygous, as would be expected from an inbred laboratory mouse strain.

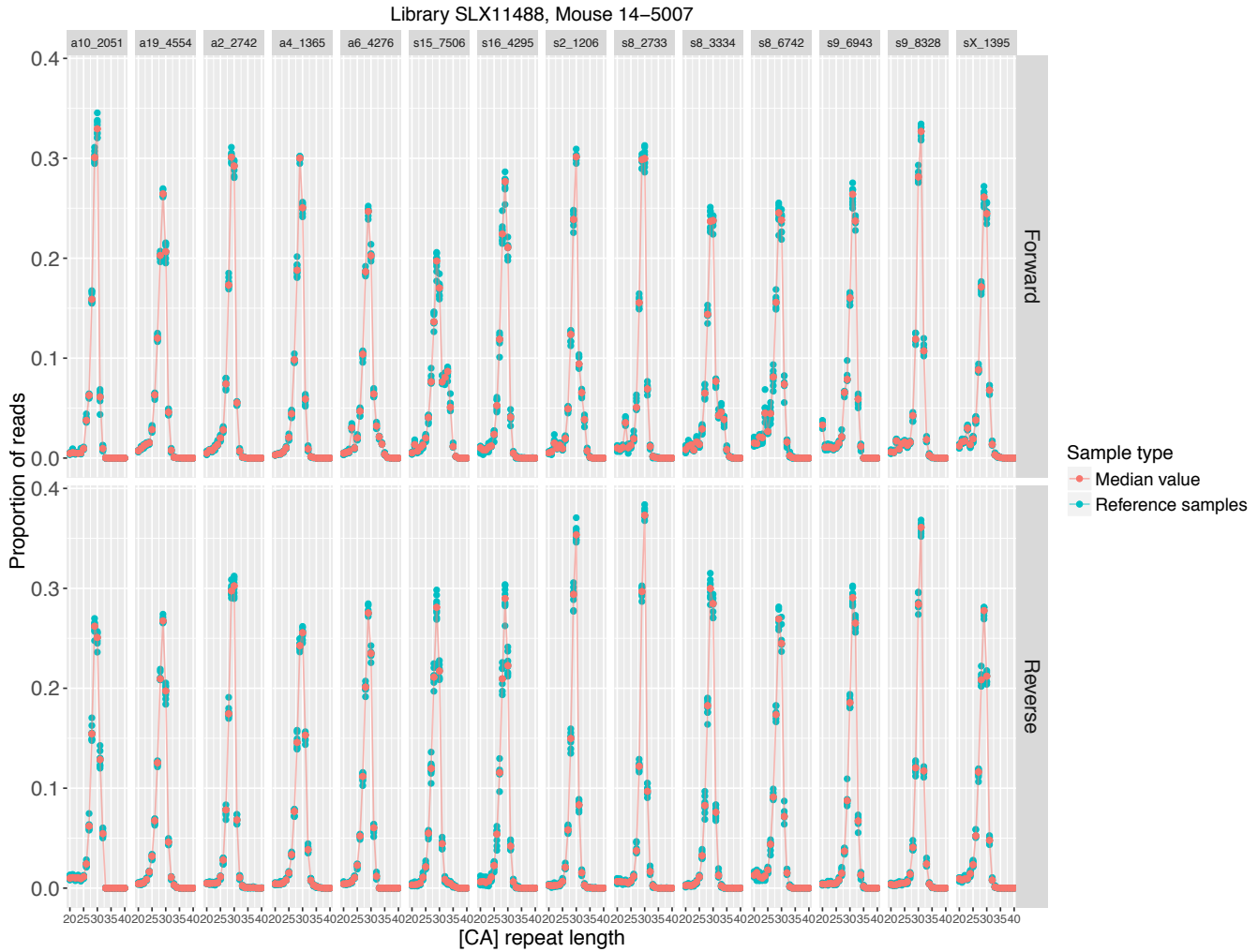

Fig. 5.1 Scatter plot with line annotation showing representative distributions produced from reference material sequencing. The pipeline for calling Φ values is only able to analyse single peaks thus loci displaying two peaks or disparity between forward and reverse reads were removed from downstream analysis. Each reference distribution was formed from 8 technical replicates of reference material obtained from tail samples diluted to the template copy equivalent of one murine crypt equivalent. Amplification and sequencing in parallel to single crypts from the same individual was then performed. In this particular example, locus s15_7506 was excluded from downstream analysis, highlighted in Figure 5.2.

5.1.2 Determining Φ value thresholds for interpretation of clone size

Before clone sizes for wild-type or Msh2 deficient epithelium could be interpreted, Φ value thresholds had to be set to enable differentiation between wild-type, partly populated and wholly populated crypts. These thresholds were varied on a locus by locus basis. Using the data produced from sequencing of YFP+ and YFP- crypts, described in Section 4.4, the spread of Φ estimates for YFP+ and YFP- crypts can be obtained. In addition, using the model described by Kozar et al [54], it is also possible to simulate the predicted spread of Φ estimates within the PPC population. The spread of wild-type (YFP-), wholly populated (YFP+) and simulated partly populated crypts can be scaled to simulate the distribution of Φ estimates expected in the colon of a 300 day old mouse, Figure 5.3A. However, when the different populations are highlighted, the overlap between different populations can be seen, Figure 5.3B. As a result, the setting of thresholds will likely label a proportion of PPCs containing small clones as wild-type and another proportion of PPCs containing large clones as WPC. This must be accounted for when modelling the clone size distributions.

The proportion of PPCs will remain constant with age whilst the proportion of wild-type and wholly populated crypts varies with age. It is therefore necessary to capture the entire wild-type and WPC population within each threshold or the PPC population will fluctuate in size as a result of age related change in wild-type and WPC incidence. Due to overlap of the PPC Φ distribution and the wild-type Φ distribution, the setting of a threshold that includes all of the wild-type crypts will lead to a fraction of the PPC population being labelled as wild-type. The proportion of PPCs labelled as wild-type will be far higher than the proportion of PPCs labelled as WPC, Figure 5.4. This will need to be accounted for when modelling the expected outcomes from crypt sequencing.

In wild-type epithelium, the majority of crypts are expected to have the wild-type length of microsatellite due to a relatively low mutation rate. In Msh2 deficient epithelium, this mutation rate is expected to be much higher with a concomitant increase in the clone incidence. This can be observed when the Φ value distribution of crypts isolated from wild-type mice and Msh2 deficient mice at 28 and 70 days post-induction of Msh2 knockout, Figure 5.5. Using this plot, thresholds were determined by visual inspection of the peak closest to $\Phi = 0$, which was interpreted as representing the crypts with wild-type microsatellite length. Thresholds were set at the tail of these distributions thus classifying each Φ value as either wild-type or mutant.

A second threshold was required to partition mutant crypts into those which were likely

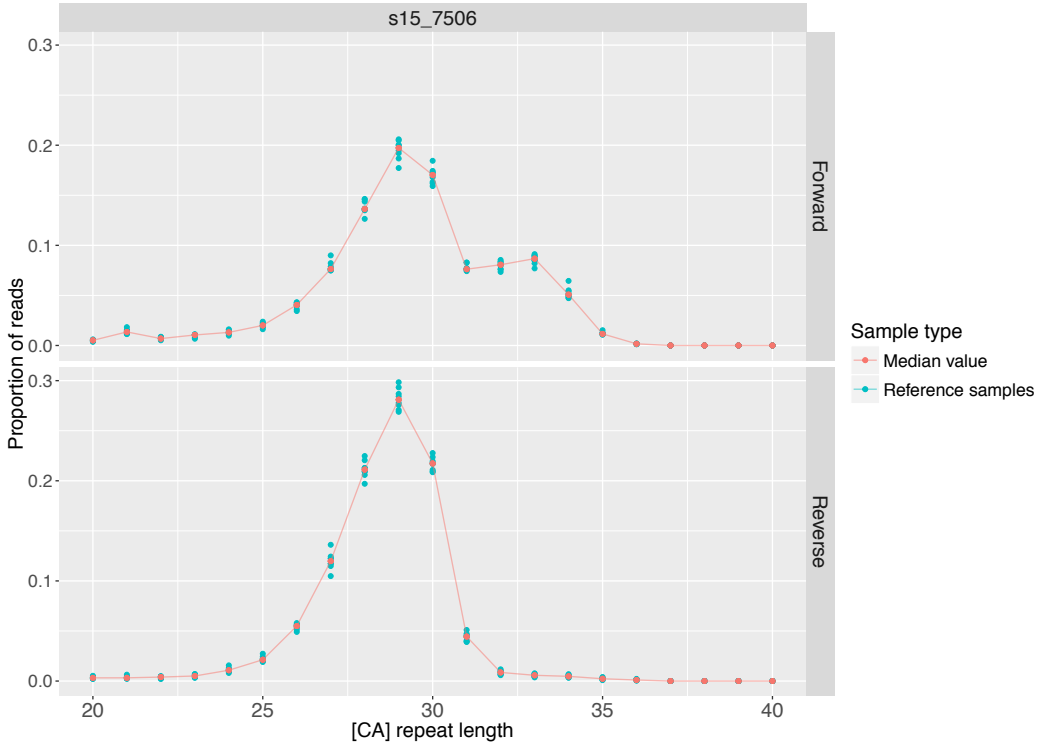

Fig. 5.2 Scatter plot with line annotation showing distribution of locus s15_7506 produced from reference material sequencing. An enlarged version of the distribution shown in Figure 5.1 illustrates the disparity in forward and reverse read distributions. As a result, this distribution was removed from downstream analysis. The same manual exclusion of bimodal distributions was performed on all loci and in all mice studied.

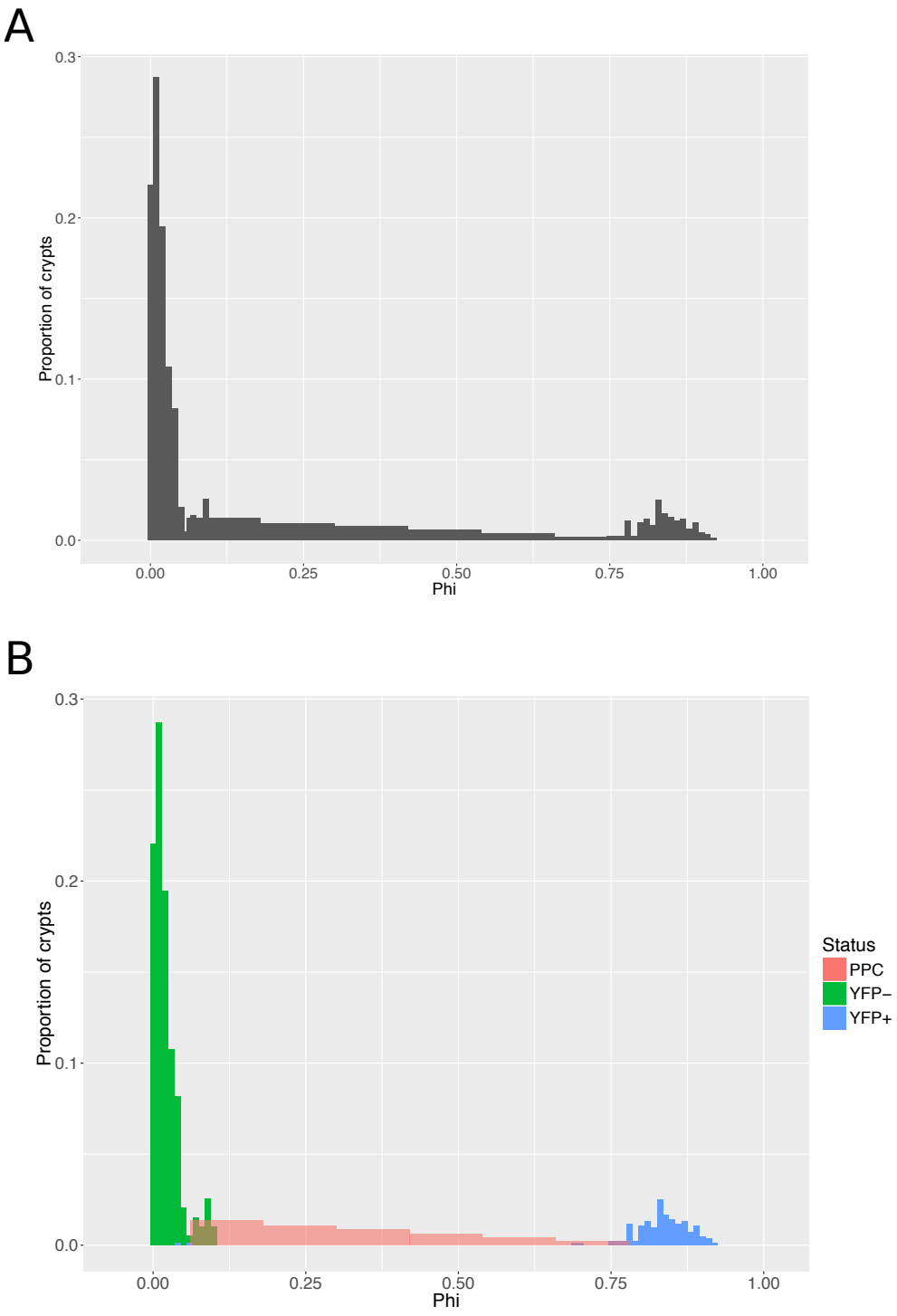

Fig. 5.3 Φ value estimate distributions for YFP+ and YFP- crypts generated from the method described in Section 4.4 and a predicted PPC population Φ value spread generated from the model described by Kozar et al [54]. The proportions of each population were scaled to represent that expected in the colon of a 300 day old mouse and the mutation rate increased 10-fold so as to highlight the PPC and WPC. A) Histogram displaying the simulated spread of Φ value estimates expected from the colon of a 300 day old mouse using actual data generated for YFP+ and YFP- crypts and theoretical estimates of Φ value spread in the PPC population. B) Histogram displaying the simulated spread of Φ value estimates with the different populations highlighted. It can be seen from this simulation that the setting of thresholds at the tail end of wild-type or WPC peaks will likely included PPCs.

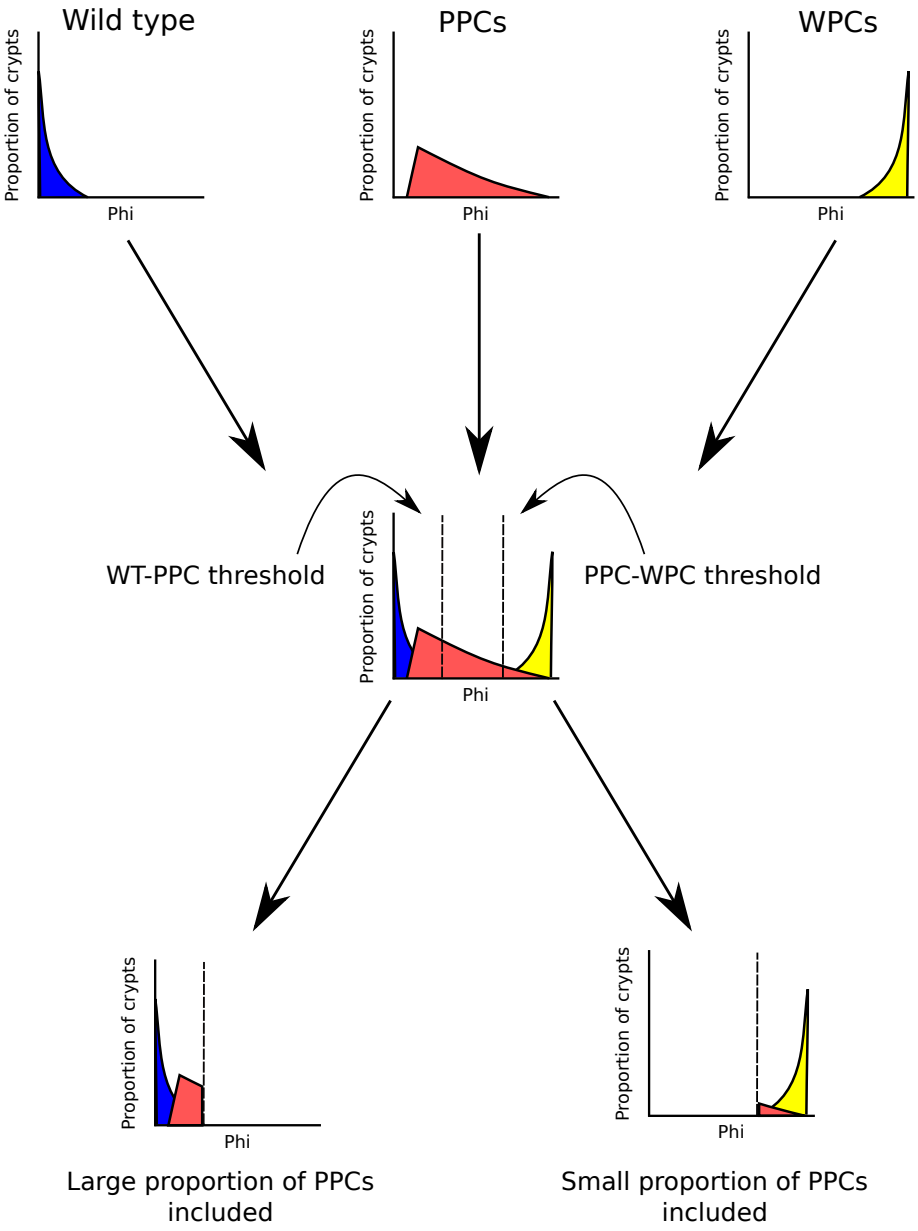

Fig. 5.4 Schematic of the use of thresholds leading to a large proportion of PPCs with small clones being labelled as wild-type and a smaller proportion of PPCs with large clones being labelled as WPC. As a result, the insensitivity of the method for identifying PPCs containing small clones needs to be accounted for in the model whilst the loss of PPCs containing large clones is less significant.

PPC or WPC. Particularly at 70 days post-induction of Msh2 knockout, there is a discernible population of wild-type crypts, PPCs and WPCs. At the majority of loci, there was a clearly discernible peak at Φ = 0.5 which was interpreted as being WPCs. The thresholds were determined by visual inspection of the WPC peak and setting of a threshold at the tail of this distribution. The threshold for differentiating between PPC and WPC can be seen in Figure 5.6. Table 5.1 summarises the final Φ value thresholds set for all 15 loci.

Setting of thresholds relied upon the presence of a distinct peak at close to Φ = 0 formed by wild-type crypts and another peak at Φ = 0.5 for biallelic loci or at Φ = 1 for monoallelic loci formed by WPCs. This was possible for most loci though some thresholds were approximated and are conservatively set to minimise identification of PPCs as wild-type or WPCs. Therefore, if either of the wild-type or WPC peaks were not clearly discernible, the loci were removed from the majority of analyses described in this chapter. The only analyses where these loci were included related to the analysis of mutable loci, Section 5.5, and analysis of mutation spectra, Section 5.6. Out of the 14 loci present in the multiplex group, 5 were excluded based on the absence of distinct wild-type and WPC peaks (loci s9_6943, a4_1365, s9_8328, a6_4276 and s15_7506). The other 9 loci were taken forward for further analysis in wild-type. The same 9 were analysed in Msh2 deficient epithelium along with the transgenic $[CA]_{30}$ which was also present.

Locus a19_4554 showed highly variable Φ value distributions at different time points post-induction of Msh2 loss and a notable lack of a WPC peak albeit for 70 days post-induction, Figure 5.7. Notably, the amount of mutant crypts in the wild-type mice (0 days post-induction) is higher than that seen in the mouse 28 days post-induction. The reason for the variability at this locus is unknown and was removed from downstream analysis. Thus, including the loci previously excluded, a total of 6 loci were removed from downstream analysis leaving 8 loci for analysis plus the transgenic $[CA]_{30}$ locus also present in the conditional Msh2 knockout mice.

The current method of estimating mutant distributions using a shifted reference distribution could be superseded by a method of using mutant reference distributions produced from crypt sequencing. Locus a2_2742 at 70 days post-induction of Msh2 knockout had a particularly large proportion of crypts with an inferred *shift* of -1 (81 out of 210 crypts tested) and, of those many had drifted to be a WPC (28 out of 81). When the distributions within these crypts are pooled and a median value plotted, the comparison of a wild-type reference distribution and a mutant reference distribution with a *shift* = -1 and a median Φ = 0.5 can

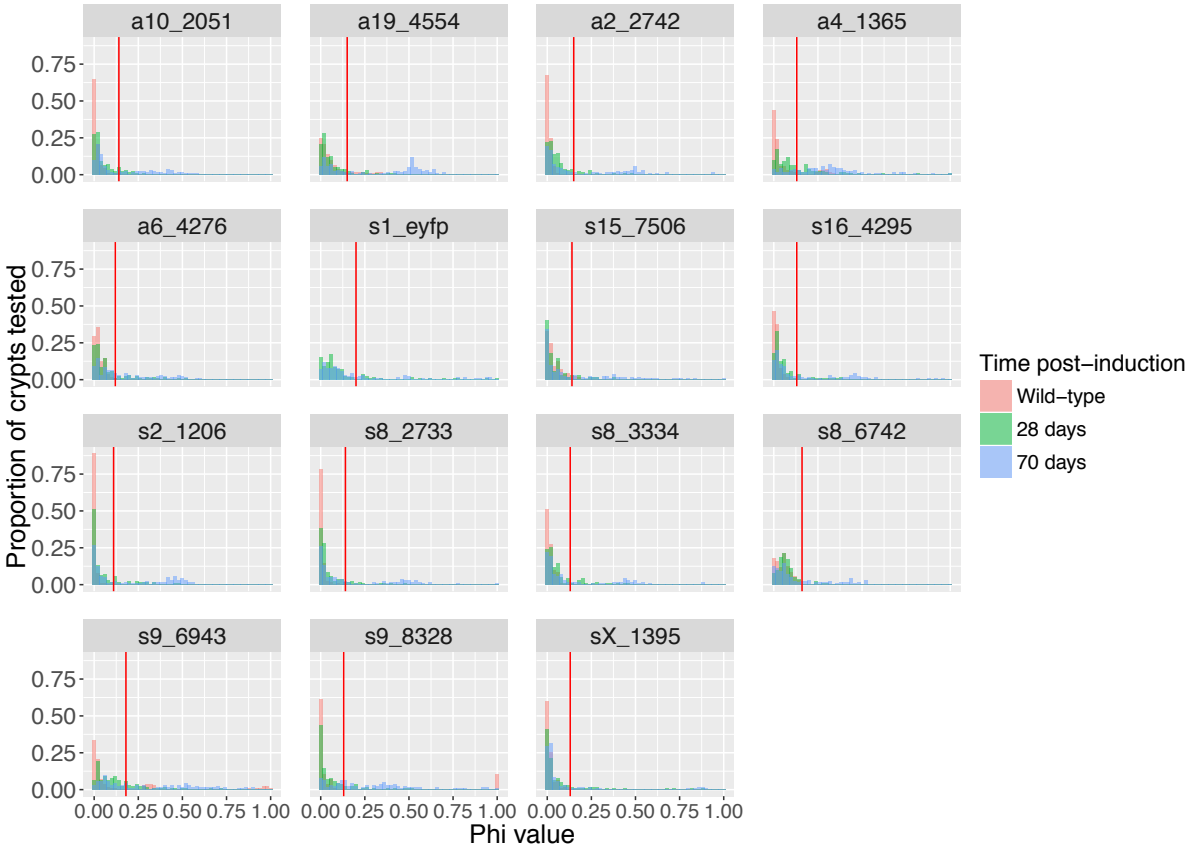

Fig. 5.5 Histograms showing frequency of Φ values at different loci in crypts isolated from wild-type and Msh2 deficient mice 28 and 70 days post-induction. The red line shows the threshold used for differentiating between wild-type and mutant crypts.

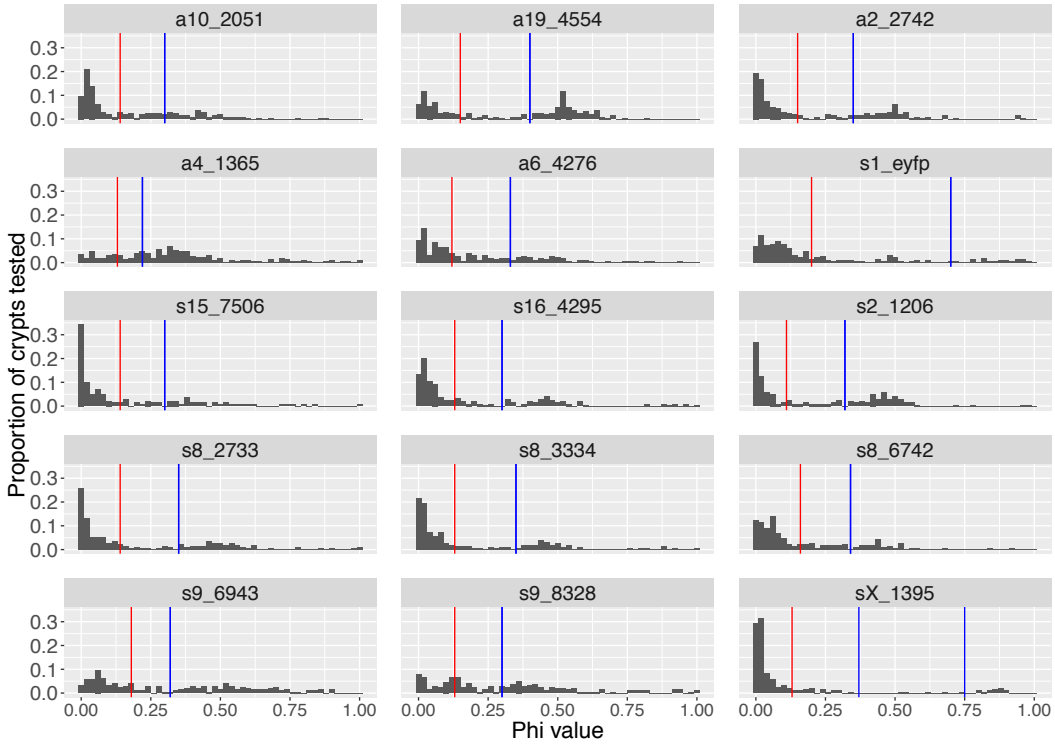

Fig. 5.6 Histograms showing frequency of Φ values at different loci in crypts isolated from an Msh2 deficient mouse 70 days post-induction. The red line shows the threshold used for differentiating between wild-type and mutant crypts. The blue line indicates the threshold used to differentiate mutant crypts as being PPC or WPC. In X-linked loci, two blue lines are present to account for monoallelic and biallelic presence in males and females respectively. Out of the 14 loci amplified in the multiplex group, 5 loci were excluded due to the absence of distinct wild-type and WPC peaks, these loci were: s9_6943, a4_1365, s9_8328, a6_4276 and s15_7506 and were removed from all downstream analyses.

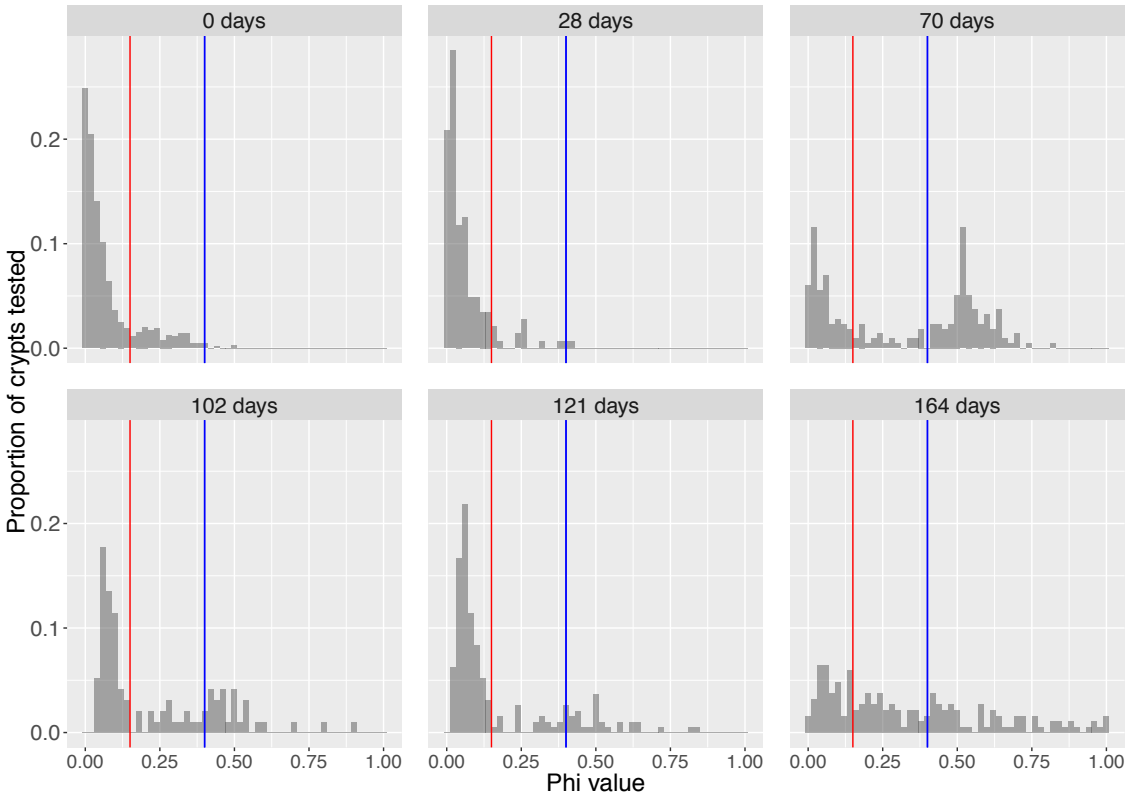

Fig. 5.7 Histograms showing the distribution of Φ values post-induction of Msh2 knockout at locus a19_4554. The proportion of mutant crypts is highly variable at different time points. The reason for this is unknown and the locus was excluded from further downstream analysis.

be observed, Figure 5.8. If, like in this example, mutant distributions could be inferred for a range of shifts at multiple loci, the need for estimating the mutant distribution would no longer be required and allow for more accurate calling of Φ. This would ultimately enable better estimates of clone size and the ability to set tighter thresholds.

5.2 Identifying intra-cryptal clone size in wild-type epithelium

Initially, sequencing was performed on wild-type crypts to show that clone size estimates could be made in normal tissue. Using the observations made by Kozar et al [54], it is possible to make predictions about the incidence of PPCs and the accumulation of WPCs expected. These predictions are based on the observation of a monoallelic locus thus the mutation rate is expected to be twice as high at a biallelic locus. In addition, mutations in the transgenic $[CA]_{30}$ microsatellite are only observed if the mutation shifts the YFP reporter in-frame. Whereas sequencing allows direct observation of all mutations and does not require in-frame shifts. The exact frequency of non-in-frame mutation is unknown thus this possibility was not included in the calculations. As a result, the predictions made about clone size incidence are likely to be an underestimate. As can be seen in Table 5.2, amplifying using the M13_33 multiplex group should lead to one WPC being observed for every 5 crypts sequenced and one PPC for every 16 crypts sequenced, in the colon of a 300 day old mouse.

Mice at different ages were selected to see if the expected pattern of consistent PPC incidence and WPC accumulation with age is observed. Two mice were initially sequenced: one aged to 76 days and one aged to 735 days. Based on the same calculations used to obtain the expected incidence in Table 5.2, and therefore assuming the same mutation rate as for the transgenic $[CA]_{30}$ locus, specific values were calculated to predict the incidence of PPCs and WPCs in the colon of mice at those ages for the 8 loci surviving filtering in Section 5.1.2. These values are outlined in Table 5.3.

Locus	WT-PPC Threshold	PPC-WPC Threshold (Female)	PPC-WPC Threshold (Male)
a10_2051	0.14	0.30	0.30
a19_4554	0.15	0.40	0.40
a2_2742	0.15	0.35	0.35
a4_1365	0.13	0.22	0.22
a6_4276	0.12	0.33	0.33
s1_eyfp	0.20	0.70	0.70
s15_7506	0.14	0.30	0.30
s16_4295	0.13	0.30	0.30
s2_1206	0.11	0.32	0.32
s8_2733	0.14	0.35	0.35
s8_3334	0.13	0.35	0.35
s8_6742	0.16	0.34	0.34
s9_6943	0.18	0.32	0.32
s9_8328	0.13	0.30	0.30
sX_1395	0.13	0.37	0.75

Table 5.1 Thresholds used to differentiate between wild-type, PPC and WPC crypts based on the inferred Φ value.

Crypt Status	Incidence	1x $[CA]_{30}$	14x $[CA]_{30}$ (# in M13_33)
WPC	1680 per 10^5 crypts	1 per 60 crypts	1 per 5 crypts
PPC	458 per 10^5 crypts	1 per 218 crypts	1 per 16 crypts

Table 5.2 Estimates of the number of WPCs and PPCs observed in the M13_33 multiplex group. The estimates are adapted from Kozar et al [54] for colon of a 300 day old mouse. The estimates account for autosomal loci being biallelic but only include the rate for in-frame mutations as the rate of conversion to out of frame mutations is not known.

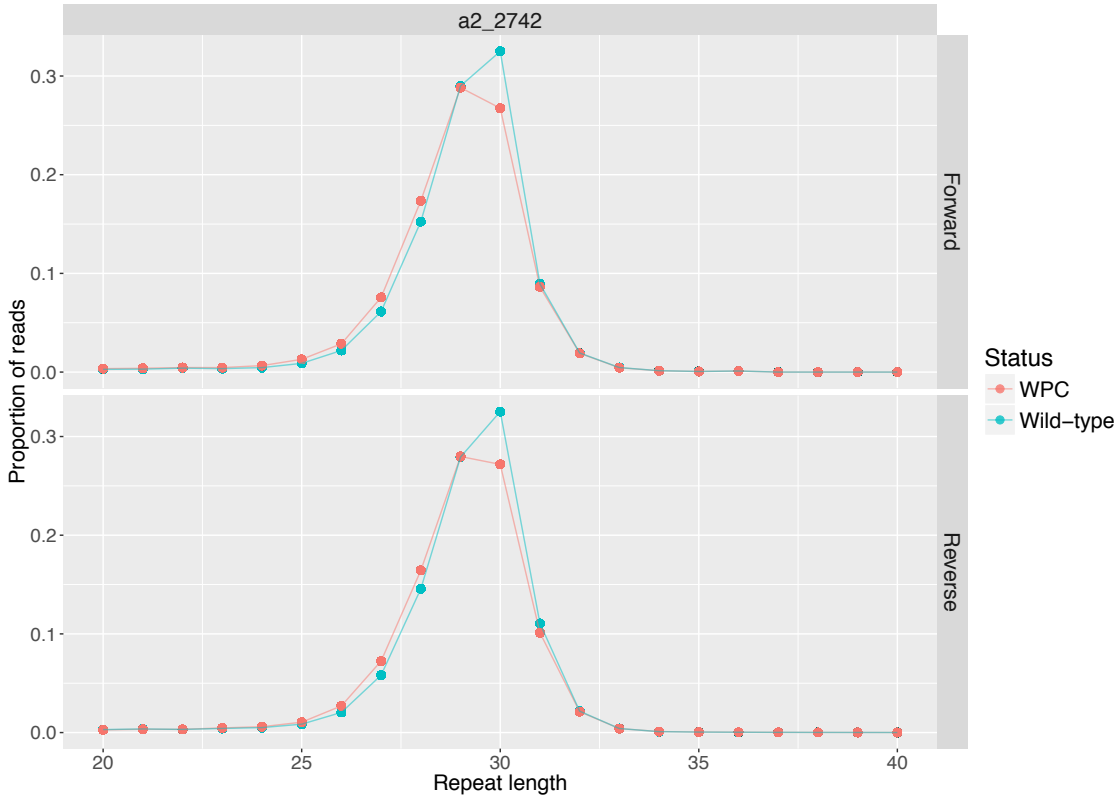

Fig. 5.8 Scatter plot showing a wild-type reference distribution at locus a2_2742 with a mutant distribution generated from crypt sequencing data. The crypts used to produce the mutant distribution were inferred as having *shift* = -1 and Φ = 0.4 to 0.6

Age (days)	**PPC**		**WPC**	
	Incidence (per 10^5 crypts)	**Expected incidence (using filtered M13_33)**	**Incidence (per 10^5 crypts)**	**Expected incidence (using filtered M13_33)**
76	458	1 per 27 crypts	425	1 per 29 crypts
735	458	1 per 27 crypts	4116	1 per 4 crypts

Table 5.3 Predicted WPCs and PPCs from wild-type sequencing of mice at 76 days versus 735 days from sequencing of 8 loci. The incidences take into account biallelic presence of endogenous microsatellites but only accounts for in-frame mutations thus represent a conservative estimate.

Sequencing of microsatellites in crypts from these two mice revealed incidences of PPCs and WPCs lower than that predicted from inferences made from the Kozar et al dataset, Figure 5.9. A mean percentage of 3.3% of crypts were PPCs, equivalent to 1 in 30 crypts, compared with the expected 1 PPC per 27 crypts. In the 76 day old mouse, 1.3% of crypts were WPCs, equivalent to 1 in 71 crypts, compared with the expected 1 WPC per 29 crypts. In the 735 day old mouse, 5.7% of crypts were WPCs, equivalent to 1 in 17 crypts, compared with the expected 1 WPC per 4 crypts. This lower than expected clone incidence is indicative of a mutation rate at endogenous loci lower than that observed at the transgenic $[CA]_{30}$ locus.

However, the clear WPC accumulation with age seen in these wild-type mice is precisely what would be expected. A 2.1-fold increase in PPC incidence was observed between the two mice. A 6.1-fold range of PPC incidence was observed in the Kozar et al [54] murine colon dataset (range: 74.2 to 452.5 per 100,000) thus a 2.1-fold difference is well within the expected range such that the increase in PPC incidence with age is likely due to sampling error alone. This experiment serves as proof of principle for the detection and interpretation of intra-cryptal clone size variation based on microsatellite sequencing data alone and suggests mutation rate heterogeneity between the transgenic and endogenous $[CA]_{30}$ loci.

To enable observation of age related changes in clone size incidence in mice across all loci, the same analysis was applied to mice with Msh2 deficient epithelium.

5.3 Identifying and interpreting intra-cryptal clone size in Msh2 deficient epithelium

The expected incidence of clones in epithelium deficient in mismatch repair pathways is not known but will be significantly higher than that seen in wild-type epithelium. However, the same pattern of constant PPC incidence and WPC accumulation with age is expected, as would be seen in wild-type epithelium. Furthermore, as values for the functional stem cell number and stem cell replacement rate within the mouse colon are known, the expected clone size distributions can be predicted, using the mathematical model described by Kozar et al [54]. By combining observations of intra-cryptal clone size variation in Msh2 deficient epithelium with simulated predictions of clone size distributions, it will be possible to validate microsatellite sequencing as a viable means of observing expected intra-cryptal clone size variation.

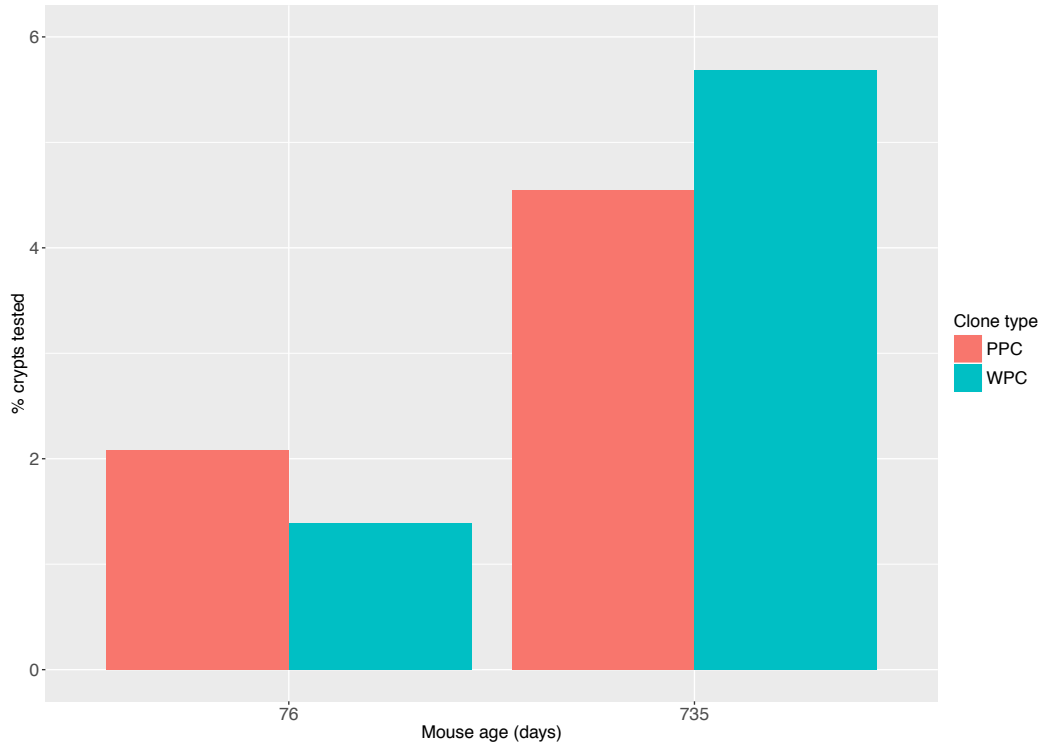

Fig. 5.9 Histogram showing the number of PPCs and WPCs observed in a 76 day old mouse (144 crypts tested) compared with a 735 day old mouse (176 crypts tested). The number of PPCs remain relatively constant with age (2.1% of crypts tested in the 76 day old mouse versus 4.5% of crypts tested in the 735 day old mouse, a 2.1-fold increase) whilst a 4.4-fold increase in WPCs is observed (1.3% of crypts tested in the 76 day old mouse versus 5.7% of crypts tested in the 735 day old mouse).

Mice with a tamoxifen inducible Cre recombinase expressed from the intestinal epithelium specific promoter *Villin* were induced leading to recombination and loss of function at the *Msh2$^{fl/fl}$* locus, described in Sections 2.1.3 and 2.1.4. Following induction, mice were taken at varying time points (range = 28 days to 164 days) and single crypts were isolated and microsatellite sequencing performed. As can be seen in Figure 5.10, across all loci, a robust increase in the number of WPCs is observed and a relatively consistent frequency of PPCs is seen at all time points post-induction of Msh2 knockout. It should be noted that one of the 8 endogenous loci was removed from this analysis as only two time points were available due to low read depth. However, the Msh2 knockout mice were also crossed onto the Rosa26-$[CA]_{30}$-eYFP strain thus the transgenic $[CA]_{30}$ locus was also analysed in this cohort. When an average PPC and WPC frequency across all loci is calculated, a pattern of clone size dynamics expected from continuous labelling is observed, Figure 5.11. These observations show that microsatellite sequencing and subsequent analysis can be used to observe and interpret clone size distributions to produce a pattern of age related clone size change close to what would be expected.

To confirm that the pattern of clone size distributions observed in Msh2 deficient epithelium fits with what would be expected from clone size distributions in murine colonic tissue, simulations of clone sizes were performed using the model described in the Kozar et al study [54]. In wild-type mice, the mutation rate at the transgenic $[CA]_{30}$ locus is calculated to be $1.1x10^{-4}$ mutation events per cell mitosis. As the loss of Msh2 is expected to lead to a significant rise in microsatellite mutation rate, the incidence of PPCs and WPCs is expected to rise. By adjusting the mutation rate input into the model, it is possible to simulate clone size distributions within mouse colon at variable mutation rates. The output of these simulations are shown in Figure 5.12.

Comparison of the simulated clone size distributions with the clone size distributions observed from microsatellite sequencing shows a close similarity with simulated clone size distributions at a 175-fold higher mutation rate, Figure 5.13. The accumulation of WPCs closely matches the simulated data suggesting that the threshold for differentiating between large clones and WPCs is effective. The frequency of PPCs is slightly lower than what would be predicted by the model and indicates that the threshold for differentiating between wild-type from PPCs is overly conservative, as expected and discussed in Section 5.1.2. Using the simulated PPC clone size distribution, Figure 5.14, it can be seen that the majority of clones are 1/7 and 2/7 in size (equivalent to a Φ value of 0.14 and 0.28 respectively, at a

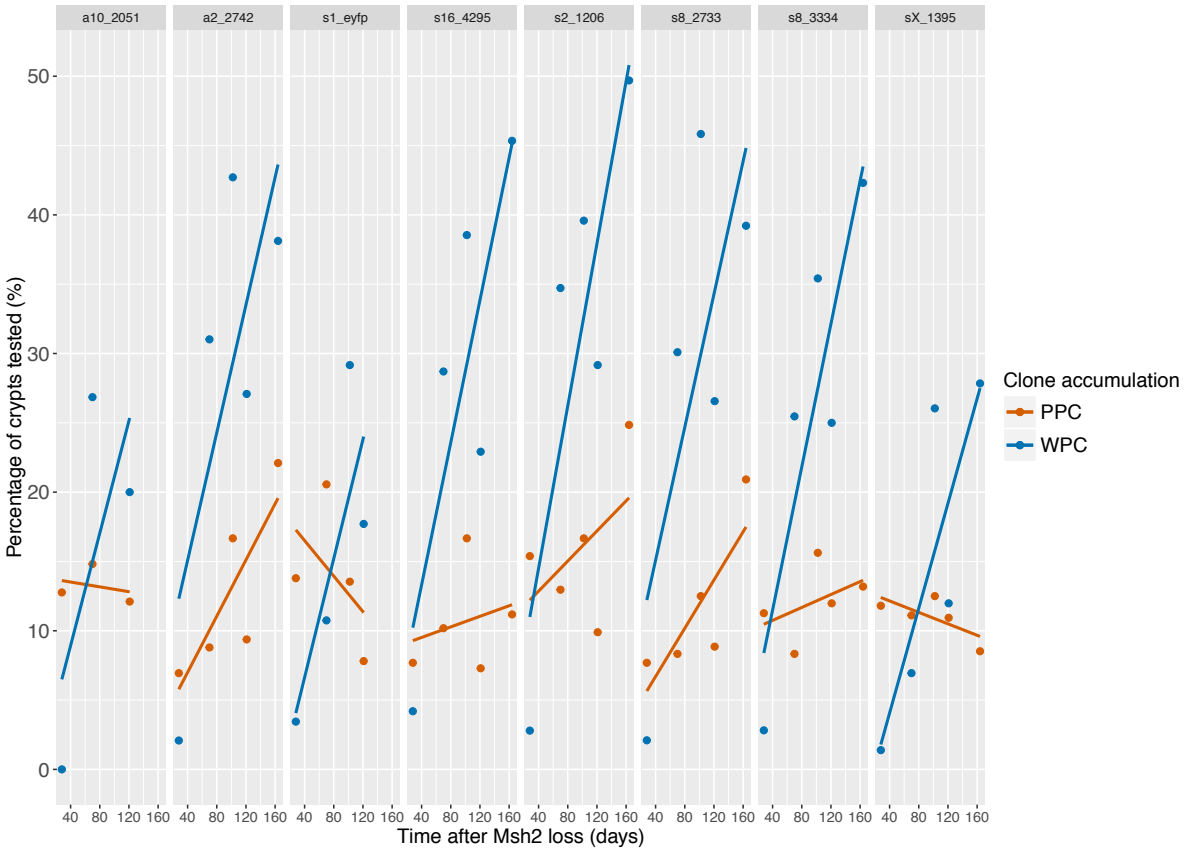

Fig. 5.10 Scatter plot showing relative changes in PPC and WPC percentages after induction of Msh2 knockout in intestinal tissues at each loci tested. Across all loci, there is a trend towards PPC percentage remaining relatively constant with a distinct increase in the percentage of WPC detected at each locus. One male mouse was analysed at each of the following time points: 28, 70, 102, 121 and 164 days post-induction with a total of 141, 210, 96, 164 and 153 crypts analysed at each time point respectively.

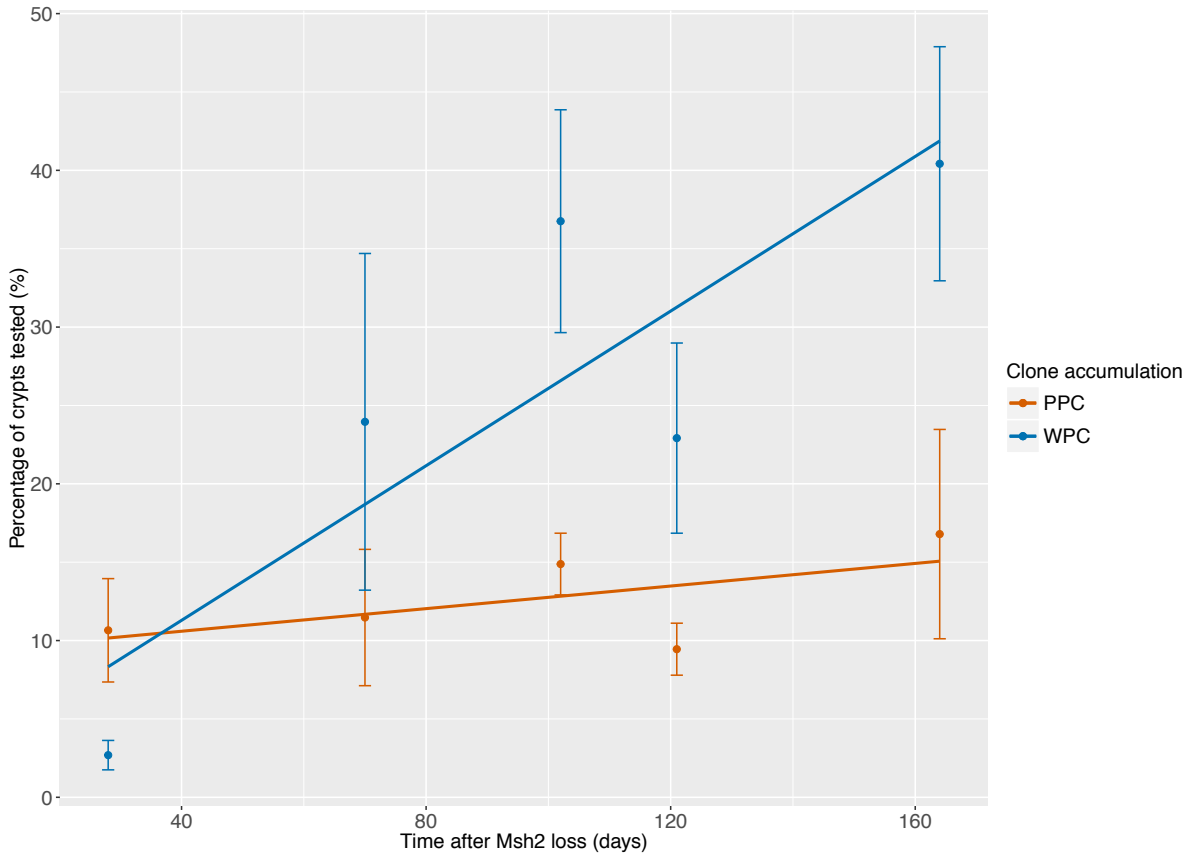

Fig. 5.11 Scatter plot showing relative changes in PPC and WPC percentages after induction of Msh2 knockout in intestinal tissues. The values were calculated by averaging across all loci tested with error bars depicting the standard deviation of estimates between loci at each time point.

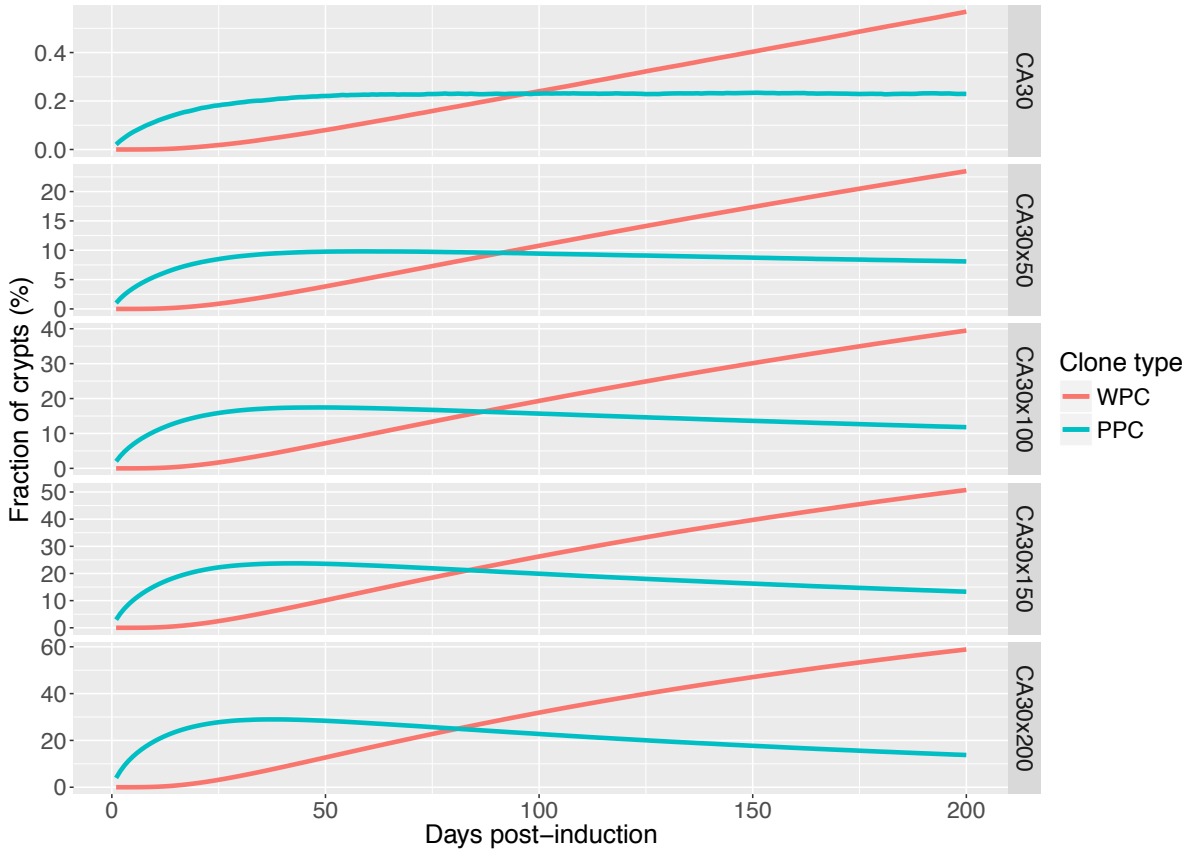

Fig. 5.12 Smoothed distributions of PPC and WPC dynamics based on varying mutation rates. These simulations were produced from the same model used by Kozar et al [54], with varying input of mutation rate.

monoallelic locus). The expected Φ estimate for clone sizes of 1/7 and 2/7 at a biallelic locus would be 0.07 and 0.14 respectively. The average threshold value for differentiating between wild-type and PPC across the analysed loci is 0.143, therefore, it is highly likely that the thresholds set will miss any clone of 2/7 in size or smaller. When this is accounted for in the model, the data now fits well with the expected clone distribution, Figure 5.15. This indicates that microsatellite sequencing is insensitive to small clone detection thus requiring adjustments to the model to account for this. Nonetheless, with this adjustment, the observed data closely matches the simulated data and indicates that the method is particularly effective at identifying WPCs.

5.4 Observation of WPC accumulation can be used to infer microsatellite mutation rate in Msh2 deficient epithelium

Using the model developed by Kozar et al [54], it is possible to infer the functional stem cell number and stem cell replacement rate so long as the microsatellite mutation rate is known. Using this model, Kozar et al inferred the number of functional stem cells per murine colonic crypt to be 7 and the stem cell replacement rate in the murine colon to be 0.3 per day. As these values have already been defined for the mouse colon [54], it is instead possible to take these values and reverse the calculation to infer the microsatellite mutation rate. This inference can be done by simulating different mutation rates to find the simulated dataset that best fits our observations. Given that the PPC incidence is likely to be underestimated, as discussed in Section 5.3, the WPC accumulation rate is likely to be a more accurate indicator of the clone size distribution. As can be seen from Figure 5.16 and described in Section 5.3, the simulated data that best describes the WPC accumulation rate seen in Msh2 deficient epithelium has a mutation rate approximately 175-fold higher than that seen in wild-type epithelium. This equates to an average mutation rate across all loci of 1.93×10^{-2} mutations per mitosis or 9.63×10^{-3} mutations per allele per mitosis. To the best of my knowledge, this is the first approximation of microsatellite mutation rate in Msh2 deficient epithelium.

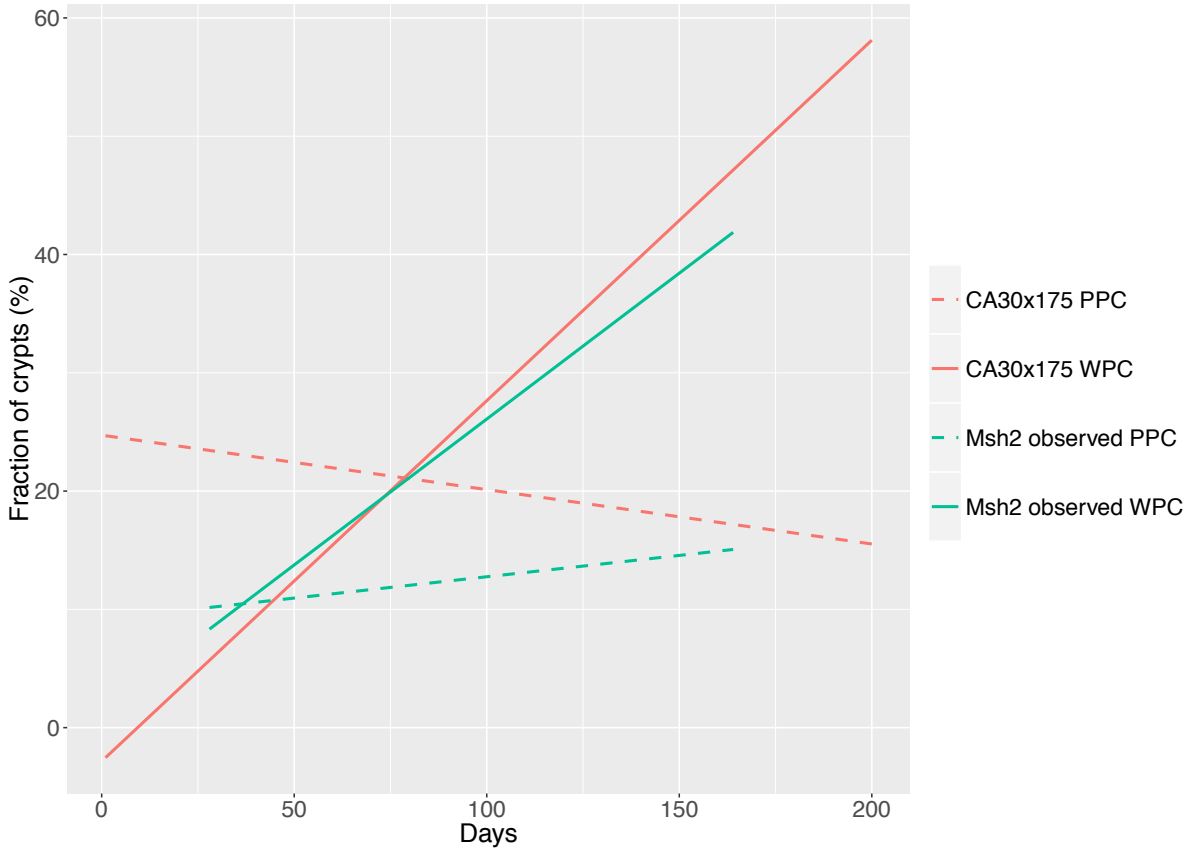

Fig. 5.13 Line graph depicting simulated clone dynamics based upon mutation rates 175-fold higher than that seen in wild-type epithelium. The WPC accumulation seen in Msh2 deficient epithelium closely matches the dynamics observed when the mutation rate is 175-fold higher than in wild-type epithelium. The PPC incidence observed from microsatellite sequencing is lower than predicted.

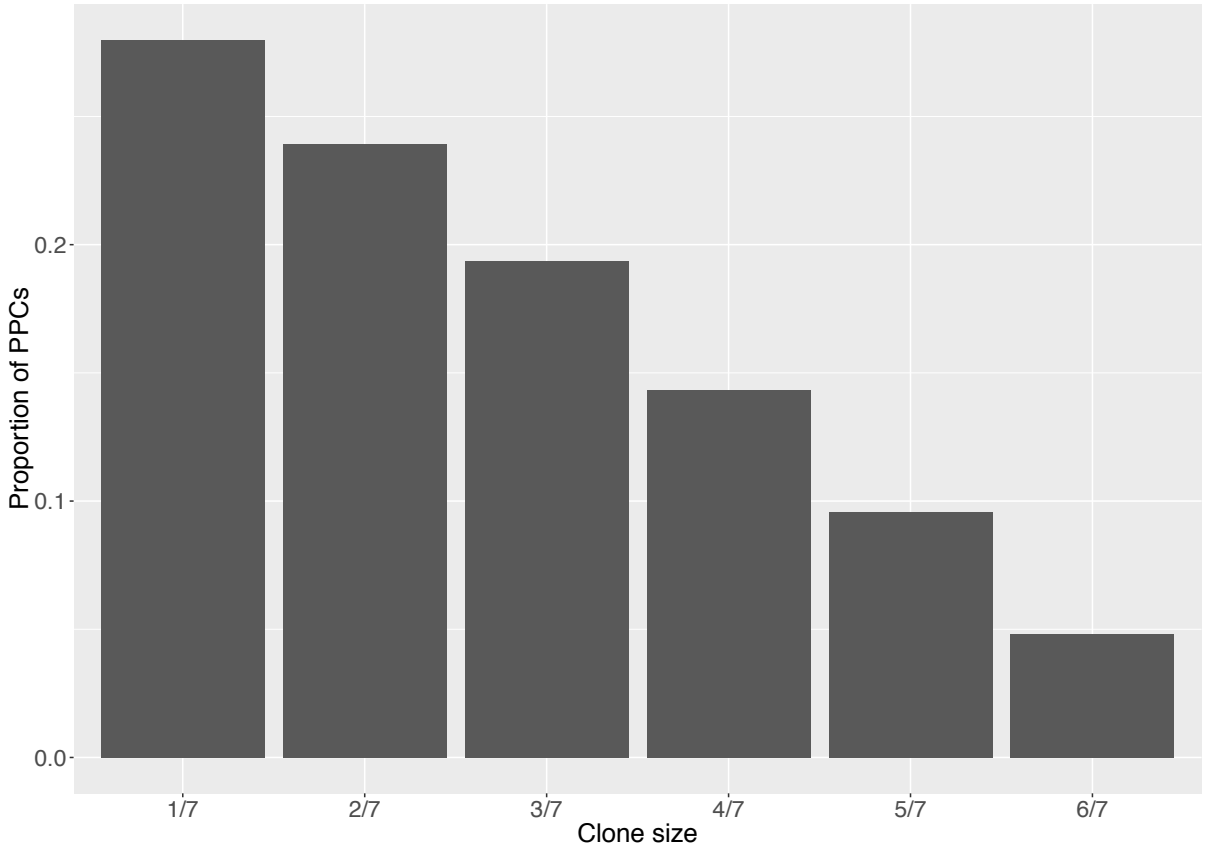

Fig. 5.14 Histogram displaying the breakdown of clone size in the simulated PPC population. These estimates were produced from the model described by Kozar et al [54].

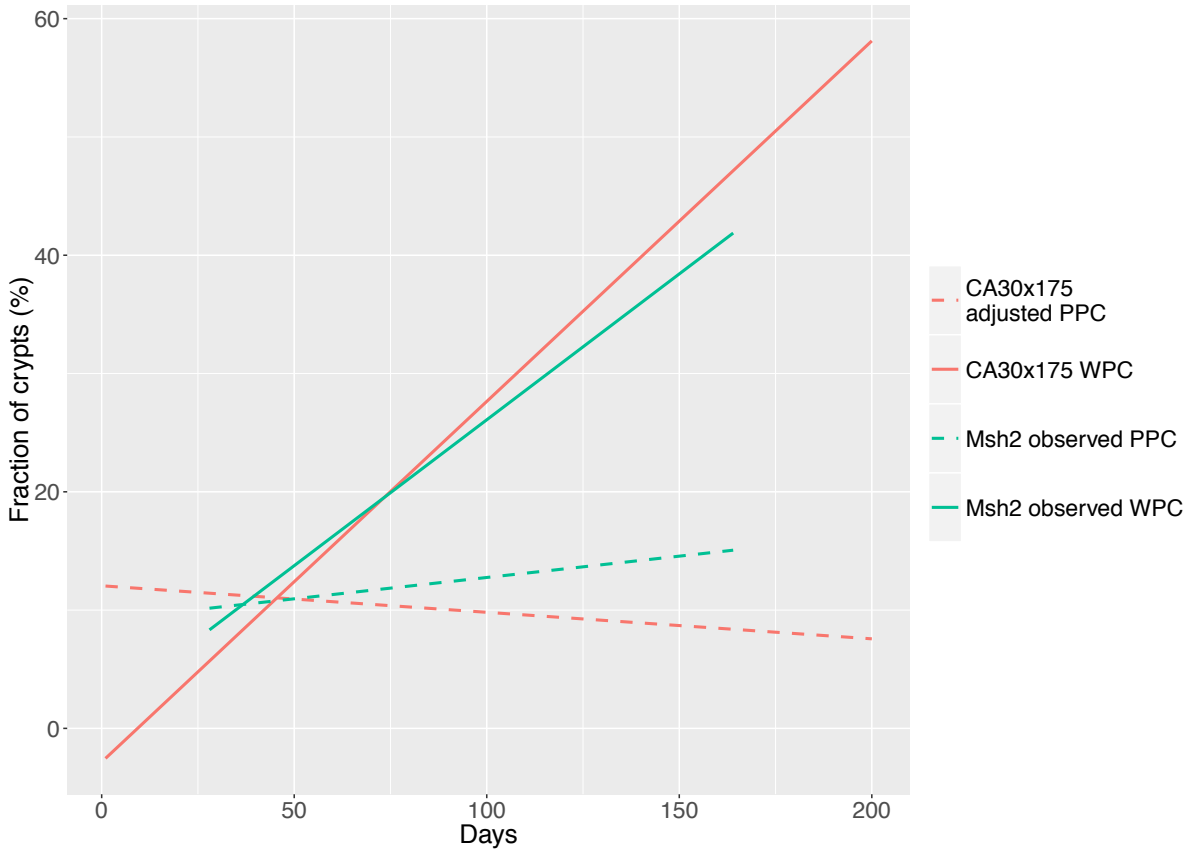

Fig. 5.15 Line graph depicting simulated clone dynamics based upon mutation rates 175-fold higher than that seen in wild-type epithelium and adjusted for small clone detection insensitivity. The clone distribution seen in Msh2 deficient epithelium closely matches the dynamics observed when the mutation rate is 175-fold higher than in wild-type epithelium and adjusted for small clone detection insensitivity predicted from microsatellite sequencing.

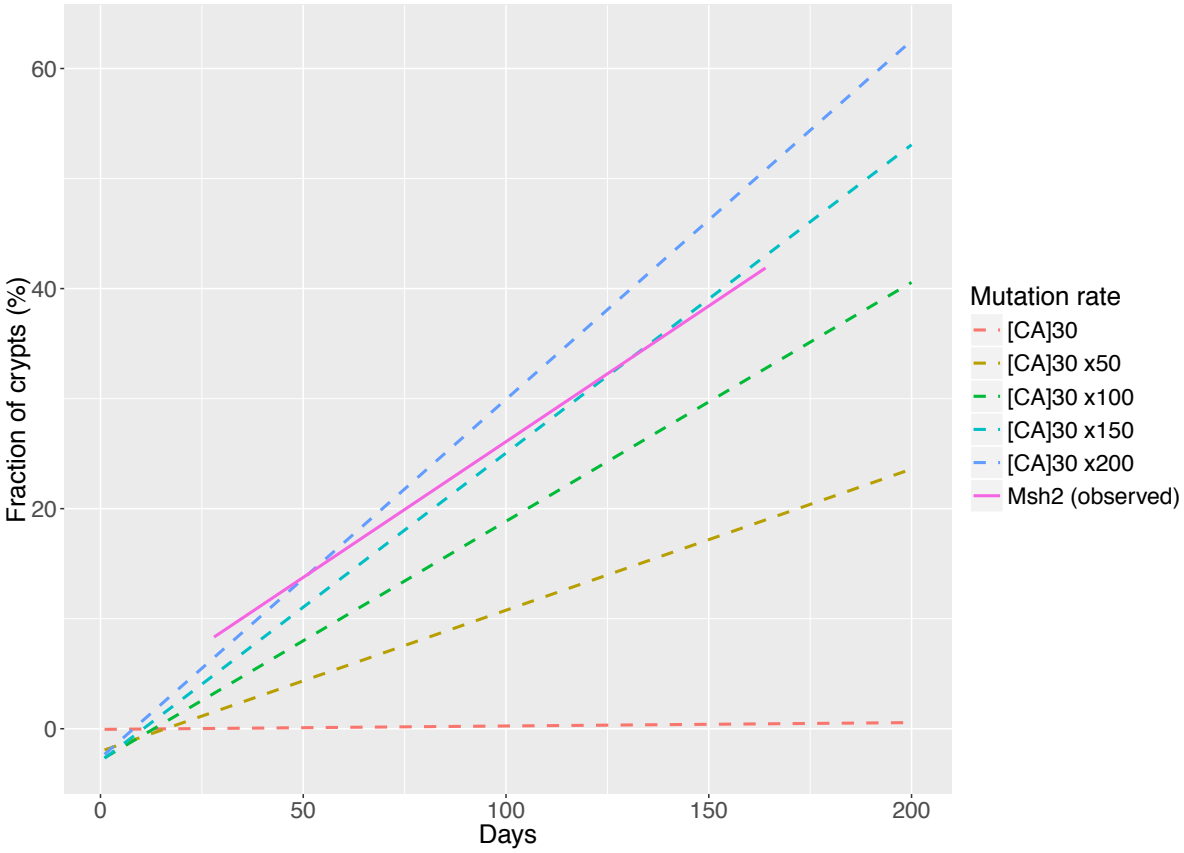

Fig. 5.16 Line graph depicting simulated WPC accumulation at different mutation rates compared with Msh2 deficient epithelium. A 175-fold mutation rate increase has the closest association with the observed WPC accumulation rate.

5.5 Germline variability predicts somatic microsatellite mutability in the mouse colon

Five loci were previously identified as being germline variable, Section 3.6. Using the clone size estimates inferred from Msh2 deficient epithelium, the relative mutability of each locus can be inferred by comparing the total percentage of mutant crypts per locus. As can be seen in Figure 5.17, 4 of the top five mutable loci were previously identified as germline variable. Furthermore, out of the top 8 loci, 6 had to be excluded from the analysis due to a lack of distinct wild-type and WPC peaks, Section 5.1.2. The 2 remaining loci are s2_1206 and the transgenic locus, s1_eyfp. This fits with previous wild-type data, Section 5.2, suggesting that the mutation rate at endogenous loci is lower than would be expected based on observations of the transgenic locus. It must be noted that though the somatically mutable loci excluded from this study are unusable in Msh2 deficient epithelium, once accurate thresholds for these loci are set, these loci could be highly informative in wild-type epithelium.

All mice in this cohort were male and, therefore, monoallelic for X-linked microsatellites. As can be seen in Figure 5.17, the X-linked loci, sX_1395, has the lowest inferred mutation rate. Thus the high mutation rate observed at the monoallelic transgenic $[CA]_{30}$ locus is notable. As this locus is the only locus under a housekeeping promoter (*Rosa26*), it is possible that the increased mutation rate is linked to the high transcriptional activity at that locus.

5.6 Microsatellite mutational spectrum is locus specific and conserved in Msh2 deficient epithelium

It has been hypothesised that microsatellites mutate via a loop insertion-deletion mechanism. This indicates that the majority of microsatellite mutations will be small scale insertions and deletions with larger changes less likely. The mismatch repair pathway is a key effector in the repair of these small insertion-deletions. Thus, knockout of Msh2 should expose this mutational process. As can be seen in Figure 5.18, there is a large bias towards small insertions and deletions at all loci studied in both wild-type and Msh2 deficient crypts. Only locus s8_6742 displayed any large scale mutations and was curiously only observed in wild-type epithelium. These observations support a mechanism of mutation leading to

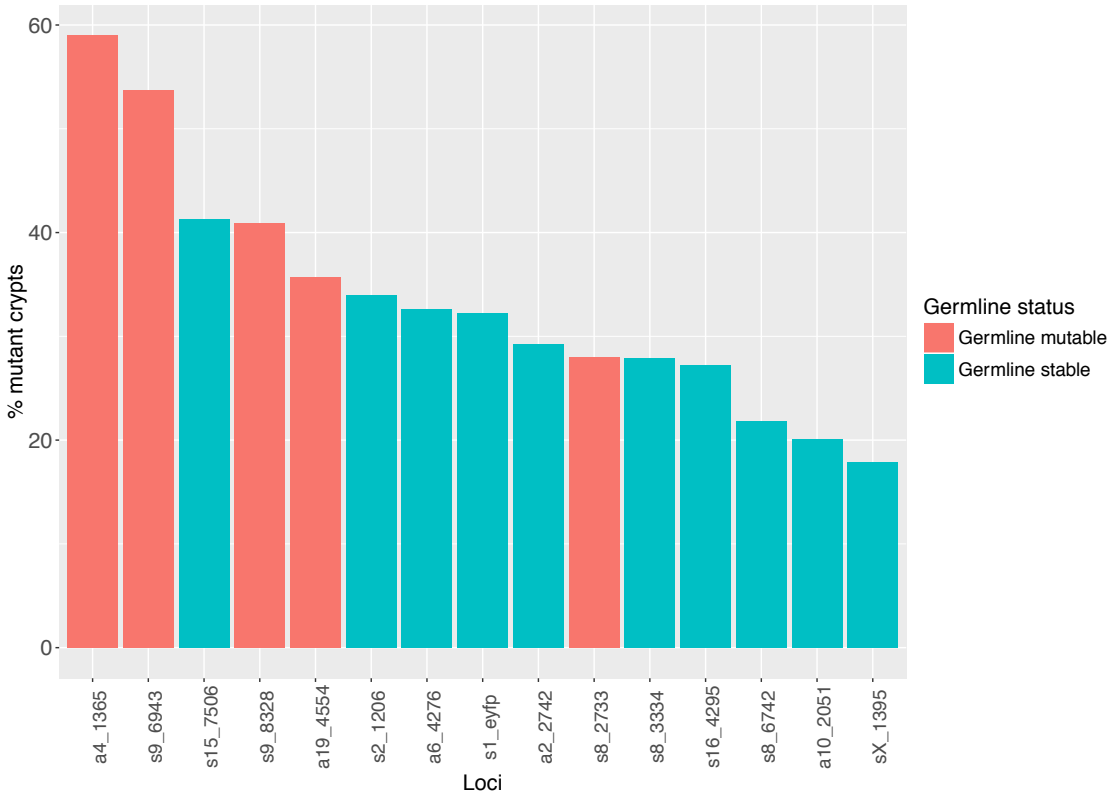

Fig. 5.17 Histogram showing ranked loci based on percentage of mutant crypts in Msh2 deficient tissue. The frequency of mutant crypts is directly proportional to the mutation rate at those loci. As can be seen from this ranking, loci that were identified as being germline variable are also somatically mutable. Notably, other than locus s2_1206, all loci above the YFP transgenic locus (s1_eyfp) were excluded from the analysis. This is inline with our observations that the mutation rate is lower than expected at the endogenous loci included in our analysis when compared with the transgenic locus. This is particularly surprising when considering that the YFP locus is monoallelic; all other loci are biallelic albeit for locus sX_1395 which has the lowest mutation rate of all the loci, as would be expected from an all male cohort of mice.

small scale insertions and deletions, in line with the hypothesised loop insertion-deletion mechanism, and is the key mutational process in Msh2 deficient epithelium. All mutational shifts across all loci are shown in Figure 5.19 highlighting the largely conserved spectrum of mutation in wild-type versus Msh2 deficient epithelium. This is highly indicative of Msh2 deficiency leading to reduced repair and no association with an increase in any other mutation process.

5.7 Discussion

In this chapter, I have described the process used to determine thresholds for interpreting Φ values generated from microsatellite sequencing data. Using these thresholds, wild-type crypts were analysed and the expected pattern of intra-cryptal clone size variation was observed. In Msh2 deficient epithelium, the incidence of clones was far higher. Simulated data, based on known values for stem cell dynamics in the mouse colon, matched the WPC accumulation observed in Msh2 deficient epithelium, based on a 175-fold increased microsatellite mutation rate. Based on these observations, the average mutation rate seen at endogenous $[CA]_{30}$ microsatellites was inferred and the mutational spectrum shown to be conserved between wild-type and Msh2 deficient epithelium.

Prior to interpretation of Φ values, some loci were excluded because of the presence of bimodal reference distributions. The most likely cause of this bimodal distribution is heterozygous microsatellite length at these loci. Though sequencing or stereotyped amplification error cannot be excluded. The presence of heterozygous microsatellite length is a relatively minor phenomena in inbred laboratory lines but this flags a potential problem in applying the method to human epithelium where polymorphism is likely to be more prevalent. Therefore, it would be useful if future development of the analysis pipeline would allow for use of bimodal distributions.

Observed Φ distributions in crypts from wild-type and Msh2 deficient epithelium 28 and 70 days post-induction of Msh2 knockout were used to set locus specific thresholds. The use of all 3 datasets enabled a threshold to be determined that differentiated wild-type crypts from mutant. Crypts at 70 days post-induction of Msh2 knockout were selected based on the presence of clearly discernible peaks in Φ estimates consistent with WPC crypts (Φ = 0.5 for biallelic loci and Φ = 1 for monoallelic loci). A second threshold was then set to enable classification of mutant crypts as either PPC or WPC. For some loci, the wild-

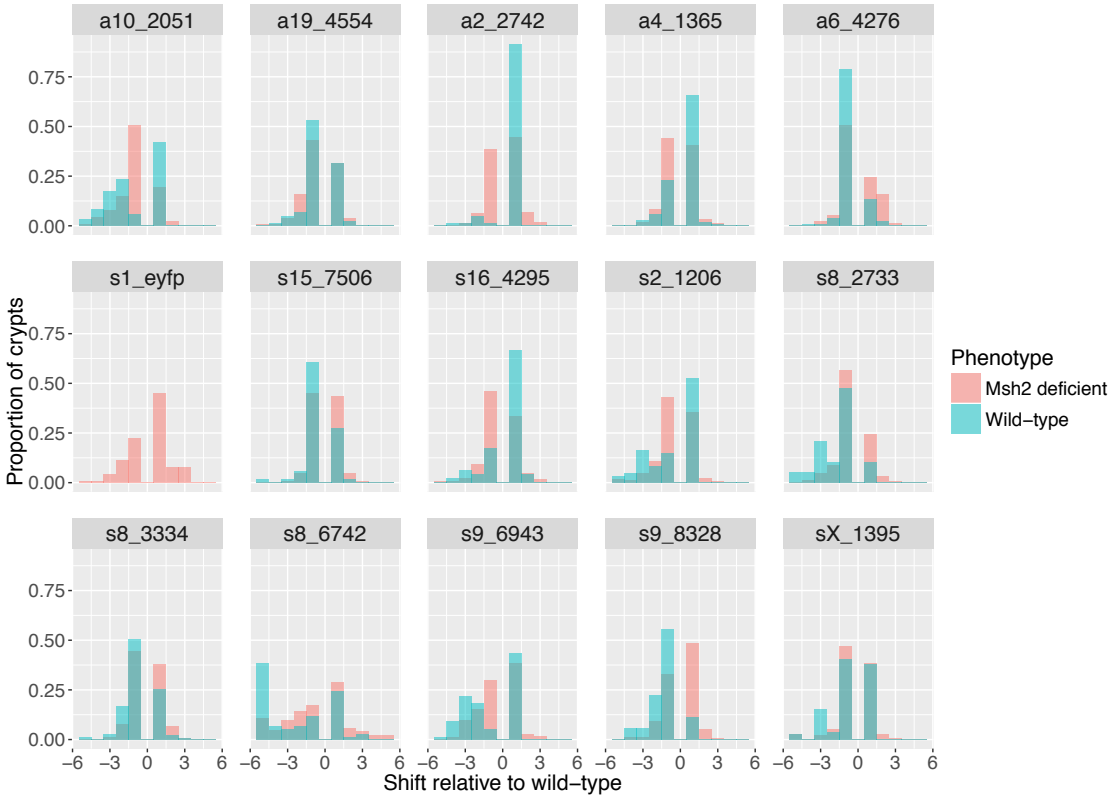

Fig. 5.18 Histograms showing range of shifts seen at different loci in wild-type crypts compared with crypts in Msh2 deficient tissue. The mutational spectra appears to be largely conserved between wild-type and Msh2 deficient epithelium.

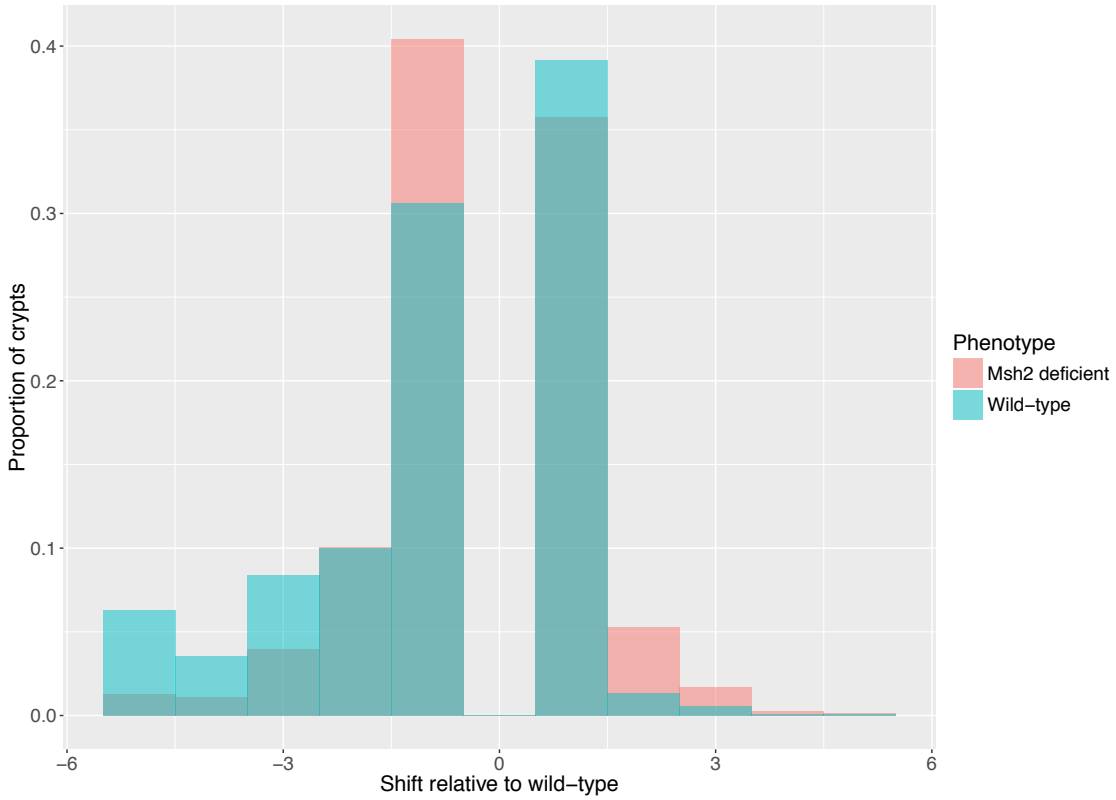

Fig. 5.19 Histogram showing the average range of shifts seen across all loci in wild-type crypts compared with crypts in Msh2 deficient tissue. The mutational spectra appears to be biased towards small scale changes, inline with a loop insertion-deletion mechanism of mutagenesis, and is relatively conserved between wild-type and Msh2 deficient epithelium.

type and WPC peaks were not clearly discernible and were removed from any analysis of clone size distributions. Additionally, locus a19_4554 was removed due to variability in the mutant fraction across different mice. The reason for this requires further study with both biological and technical reasons possible. Nonetheless, using the thresholds set, it was possible to interpret the Φ value estimated for each locus, in every crypt.

Sequencing of wild-type crypts revealed lower PPC and WPC incidences than that predicted by values calculated from the transgenic $[CA]_{30}$ locus described in the Kozar et al study [54]. This was further highlighted in Msh2 deficient crypts where the transgenic $[CA]_{30}$ locus had a far higher mutation rate compared to the majority of endogenous loci. This is particularly notable considering the monoallelic presence of the transgenic locus whilst all but one of the endogenous loci are biallelic. One possible explanation for this discrepancy is the transcriptional activity of the *Rosa26* locus predisposing the $[CA]_{30}$ in that region to an increased mutation rate.

Simulation of outcomes based on 7 functional stem cells and a replacement rate of 0.3 per day allows for predictions of age related change in clone incidence under differing microsatellite mutation rates. These simulations predicted the mutation rate in Msh2 deficient epithelium to be approximately 175-fold higher than at the transgenic $[CA]_{30}$ microsatellite. The accumulation of WPCs seemed to robustly correlate with the predicted WPC accumulation rate. However, the PPC frequency appeared to be underestimated. This disparity is likely due to a conservative threshold being set for differentiating between wild-type and partly populated crypts leading to some small clones being called as wild-type instead of partly populated. This underestimate of PPC could be accounted for in two ways: 1) the mathematical model used to predict outcomes from continuous labelling could be adjusted so that only medium and large size clones are counted such that the insensitivity to small clones is insignificant, as was shown to be an effective method in Section 5.3. 2) As more data is generated using this method, mutant reference distributions can be generated reducing the inaccuracy in Φ estimates and allowing tighter thresholds to be set, as was shown in Figure 5.8. This will increase the sensitivity of the method for detecting small clones. Furthermore, as more data becomes available, the wild-type and WPC peaks are likely to become more clearly discernible allowing for more accurate thresholds to be set. Ultimately, a combination of improved modelling and larger datasets will enable more sensitive analysis.

In addition to allowing more sensitive thresholding, the availability of more microsatel-

lite sequencing data from wild-type epithelium will allow thresholds to be set at more mutable loci that had to be excluded from this analysis. Inclusion of more mutable loci will be beneficial for two key reasons: 1) less loci will be excluded, therefore, allowing more data to be generated per crypt. 2) The more mutable loci will have a higher clone incidence thus contain more information regarding clone size distributions than the loci analysed in the dataset presented in this chapter.

Nonetheless, simulated clone outcomes with a 175-fold increase mutation rate confirmed that the clone size distribution observed in Msh2 deficient epithelium is close to what would be predicted. Furthermore, the presence of a robust correlation between simulated and observed WPC accumulation allowed for an approximation of the average microsatellite mutation rate in mismatch repair deficient epithelium, at 1.93×10^{-2} mutations per mitosis per locus or 9.63×10^{-3} mutations per mitosis per allele.

During initial sequencing of loci from reference material in different mice, 5 loci were identified as being germline variable. These loci were then included in the optimised multiplex group and analysed in both wild-type and Msh2 deficient crypts. Interestingly, these loci were also among the most somatically mutable, in Msh2 deficient epithelium, as inferred from an increased mutant crypt fraction seen at these loci. Unfortunately, in Msh2 deficient epithelium, the presence of a wild-type crypt peak was so low that setting thresholds for these loci, and many of the other mutable loci, was unfeasible. These loci were therefore removed from clone size distribution analysis. Though these loci are not useful for clone size distribution studies in the presence of mismatch repair deficiency, they could potentially be useful as sentinel loci for the detection of mismatch repair deficiency in normal epithelium or in tumour samples. It should also be noted that the lack of wild-type and WPC peaks at these germline variable loci means that the thresholds set are approximations thus the incidences of mutant crypts should not be considered absolute values and instead should be considered relative to each other.

In addition to studying the clone size distribution and mutation rate in Msh2 deficient epithelium, the spectrum of mutational shifts was also analysed. At the majority of loci, there was a large bias towards small insertion and deletion events consistent with the loop insertion-deletion mechanism of microsatellie mutation. Furthermore, this spectrum of mutation was conserved in Msh2 deficiency supporting the notion that microsatellite length change is a constant dynamic between mutation and repair that is biased towards mutation in the absence of mismatch repair pathway competency.

In this chapter, I have described the application of microsatellite sequencing to wild-type and Msh2 deficient mouse colon. These studies have shown the ability to observe and correctly interpret intra-cryptal clone size variation consistent with those expected from continuous labelling studies. Through adaptation of the mathematical model used by Kozar et al, and adjustment for only PPCs containing clones of 3/7 or above, it was shown that the clone size distributions observed in Msh2 deficient epithelium are consistent with what would be expected at microsatellite mutation rate 175-fold higher than that seen in wild-type epithelium. Furthermore, through use of the same mathematical model, the microsatellite mutation rate observed in Msh2 deficiency was inferred. Finally, the mutational spectrum observed in wild-type and Msh2 deficient epithelium was assessed showing conservation of small scale alterations in microsatellite length between wild-type and Msh2 deficient epithelium. The next step in the development of this method is to show that intra-cryptal clone size variation can be detected using microsatellite sequencing in patient material.

Chapter 6

Quantifying clone size in human crypts using microsatellite sequencing

Microsatellite sequencing of single crypts in murine colon validated the approach as a method for quantifying intra-cryptal clone size variation. In this chapter, I aim to translate the microsatellite sequencing protocol for the amplification of [CA] dinucleotide microsatellites to human crypts. The translated method will then be used to quantify clone size in patient material. The application of microsatellite sequencing to clone size quantification to patient material will pave the way for larger scale studies for quantifying stem cell dynamics in the human intestine.

The microsatellite sequencing protocol optimised for murine crypts will be translated to human material. Previous mixing of 4000 plasmid copies (template copy equivalent to a single human crypt), Section 4.3, showed a clear ability of the microsatellite sequencing protocol to identify minor mutant microsatellite species. Therefore, the increased DNA content of the human colonic crypt should facilitate translation from mouse studies. However, it was unknown as to whether sequencing of human genomic sequence containing [CA] dinucleotide repeats would be feasible, for example, due to issues associated with human specific sequence context of these regions or the prevalence of polymorphic microsatellites. In addition, optimisation of the multiplex PCR protocol was done using mouse specific primers so the effect of using human specific primers also remained to be seen.

Once the amplification of [CA] microsatellites from human crypts has been established, it will be possible to amplify single human crypts isolated from patient material and identify mutant clones. Using the same mixture modelling analysis, as described in Chapter 4, in-

ferring clone sizes will be possible. Interpretation of any given clone as being WPC or PPC will rely upon the presence of distinct wild-type and WPC peaks which would allow for the setting of thresholds for Φ value interpretation. The microsatellite mutation rate in human epithelium is expected to be similar to that seen in wild-type murine epithelium, therefore, a distinct wild-type crypt peak would be expected allowing differentiation between wild-type and mutant. The setting of a second threshold to classify mutant crypts as PPCs or WPCs is likely to present novel issues as Msh2 deficient crypts cannot be used as was done for mouse.

6.1 Adaptation of murine microsatellite sequencing protocol to human material

To translate microsatellite sequencing to human material, human specific primers needed to be designed and appropriately sized amplicons obtained from human crypt equivalents, with equal balance between amplicons within the multiplex group.

6.1.1 Design of human specific multiplex PCR primers

Initially, using the human reference genome (build hg38), 63 microsatellites were identified between $[CA]_{28}$ and $[CA]_{32}$. Using the online primer design tool, BatchPrimer, primers were designed for 58 of the loci of which 53 generated a single product at the expected length when analysed using gel electrophoresis (91% success rate). These 53 successful primer pairs were then taken forward for multiplex group design. Using the online tool, MultiPLX 2.1, six multiplex groups were recommended (five of which contained 9 primer pairs and one contained 8 primer pairs). An additional multiplex group was also tested that only contained primer pairs that amplified $[CA]_{30}$ microsatellites representing 8 out of the 10 $[CA]_{30}$ loci present in the human reference genome. Each multiplex group was tested using reference human DNA diluted to the equivalent of a single human crypt and amplified using the same multiplex PCR conditions used for mouse crypt amplification. Gel electrophoresis was used to observe product formed at the expected length with comparison of the relative primer-dimer levels compared with product formation. The groups with the greatest amount of product relative to primer-dimer were combined together to expand the number of primer pairs per group. The final multiplex group contained 21 primer pairs and

contained 8 primer pairs that amplify $[CA]_{30}$ microsatellites: this multiplex group is known as hsM13_53. Information about the primers and the related genomic information can be found in Appendix A and Appendix B respectively.

6.1.2 Sequencing of human crypt equivalents

Sequencing of human genomic DNA diluted to the template copy equivalent of one human colonic crypt was done by first amplifying using the hsM13_53 multiplex group, sequencing on the MiSeq platform before assessing the balance of read depth between amplicons. As can be seen in Figure 6.1, this multiplex group displays some variability in amplicon representation in initial sequencing. The protocol was used 3 times: each time the concentration of each primer pair in the mutliplex group was adjusted to improve the representation of each amplicon. In the final sequencing run, improved balance between amplicons is observed, Figure 6.1. The multiplex group went through fewer iterations of primer concentration optimisation than the mouse work up but was informed by primer concentration optimisation required in mouse studies. As such, it is possible to see that near equal representation of each amplicon is observed despite undergoing fewer optimisation steps.

Sequencing of patient reference material diluted to the equivalent of a single human crypt revealed consistency in read distributions between technical replicates comparable to that seen when amplifying mouse crypt equivalents, Figure 6.2. Unlike the mouse sequencing data, there was far more heterozygosity observed at these loci, as may be expected from a human population compared with an inbred laboratory mouse strain. In mice, between one and five of the 14 loci amplified had to be excluded due to heterozygosity. In humans, out of 21 loci amplified, 12 had to be excluded from one patient and 13 from the other. Furthermore, only two of the loci had an allele at the length expected from the human reference genome. The source of this disparity is likely to result from: 1) the low read depth sequencing used to build the human reference genome is prone to microsatellite length calling error and 2) divergence of these loci within the human population as a result of the increased mutation rate associated with microsatellites. The expected and observed microsatellite lengths are summarised in Table 6.1.

Overall, it would appear that the pipeline for the design and validation of the multiplex PCR protocol for the amplification and sequencing of microsatellite sequencing is robust in mouse and human. The multiplex PCR can accurately call microsatellite length from single murine crypts, shown in Chapters 3 and 5, and, as shown in this section, from human

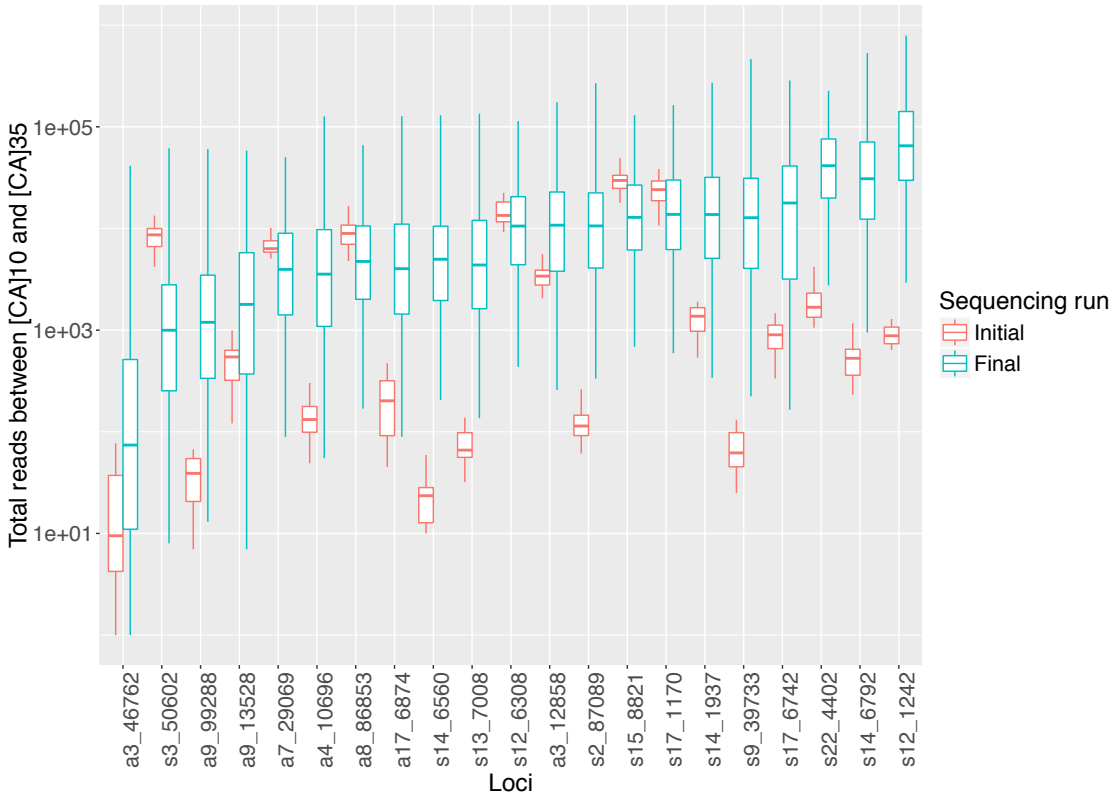

Fig. 6.1 Boxplot depicting the median read depth along with upper and lower quartiles and lowest and highest values across all amplicons in hsM13_53 used for multiplex amplification of single human crypts. The initial sequencing run was done using diluted reference material leading to highly consistent DNA template input whilst the final sequencing run was done using actual crypt material with larger variability in DNA template input. Thus explaining the reduced spread of read depth when sequencing crypt equivalent samples. However, the balance of read depth between amplicons is improved in the final sequencing run. The multiplex group went through fewer iterations of primer concentration optimisation but was informed by primer concentration optimisation done in mouse multiplex group optimisation. As such, it is possible to see that the near equal representation of each amplicon is observed despite undergoing fewer optimisation steps.

	Expected	Observed (2 alleles)	
Locus	Reference genome length	Patient 1 length	Patient 2 length
s2_87089	28	13/23	13/25
s3_50602	30	19/19	19/23
s9_39733	30	19/22	21/21
s12_1242	30	28/32	29/29
s12_6308	28	22/22	22/24
s13_7008	29	25/30	28/31
s14_1937	28	14/23	19/22
s14_6560	30	27/27	23/27
s14_6792	28	26/26	26/28
s15_8821	30	17/17	17/17
s17_6742	30	19/26	21/24
s17_1170	28	16/21	16/16
s22_4402	29	14/26	27/31
a3_46762	28	18/26	25/25
a3_12858	28	23/23	20/30
a4_10696	31	28/28	21/31
a7_29069	29	20/24	22/22
a8_86853	30	22/22	22/22
a9_13528	28	N/A	27/27
a9_99288	28	12/16	12/16
a17_6874	28	14/19	16/26

Table 6.1 Table summarising reference genome microsatellite lengths compared with the microsatellite lengths observed in two different patient samples. Only loci s14_6792 and a4_10696 out of the 21 loci observed had an allele at the expected length.

crypt equivalents. It is likely that this method could be expanded to study microsatellites within other mammalian species. The sequence context of these microsatellites is likely to be similar between mouse and human thus the use of this method in plants and other lower organisms remains to be seen but should, theoretically, be feasible.

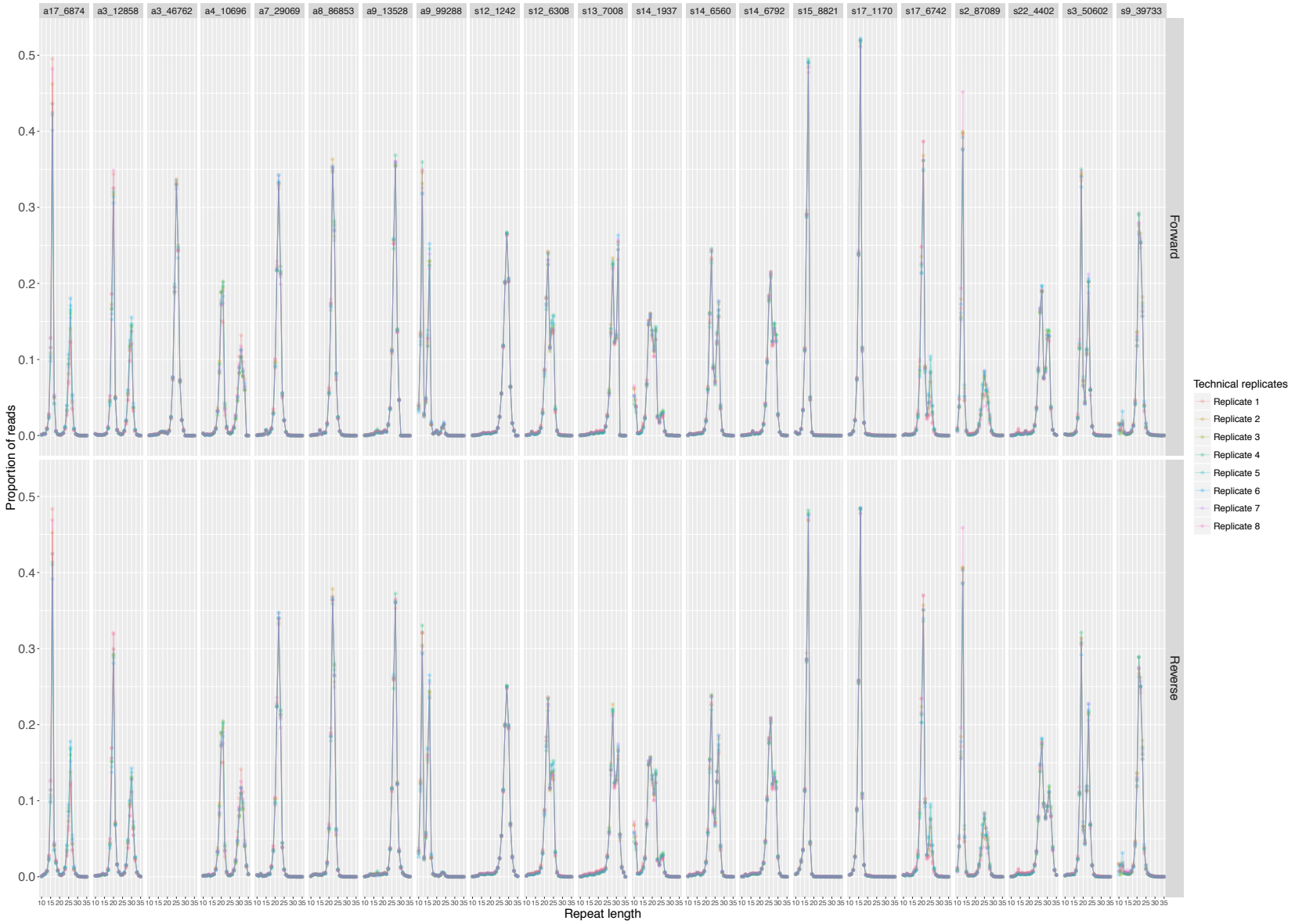

Fig. 6.2 Scatter plot with line annotation showing the read distributions of 8 technical replicates produced by amplifying reference material from one patient. Locus a3_46762 has a missing reverse read distribution due to low read depth.

6.2 Quantification of clone size from microsatellite sequencing of single human crypts

Having established that microsatellite sequencing of human crypt equivalents is possible, the feasibility of amplifying and microsatellite sequencing of actual single human crypts and subsequent quantification of clone size needed to be validated. Two patients undergoing radical colectomy for colorectal cancer donated their adjacent, normal colonic tissue to this study. These tissue samples were processed as described in Sections 2.7.2 and 2.8.4. Lysate containing single crypts were amplified using hsM13_53 and submitted for HiSeq 4000 sequencing.

The read distributions generated from sequencing of reference material, not previously shown in Figure 6.2, are shown in Figure 6.3. In this patient, there is heterozgosity of microsatellite length observed in 12 of 20 loci amplified. In the other patient sample, 13 out of 21 loci displayed heterozygosity. These loci are not amenable to mixture modelling analysis thus have to be removed from downstream analysis. After filtering for bimodal loci, the remaining distributions from the same patient displayed in Figure 6.3, are shown in Figure 6.4.

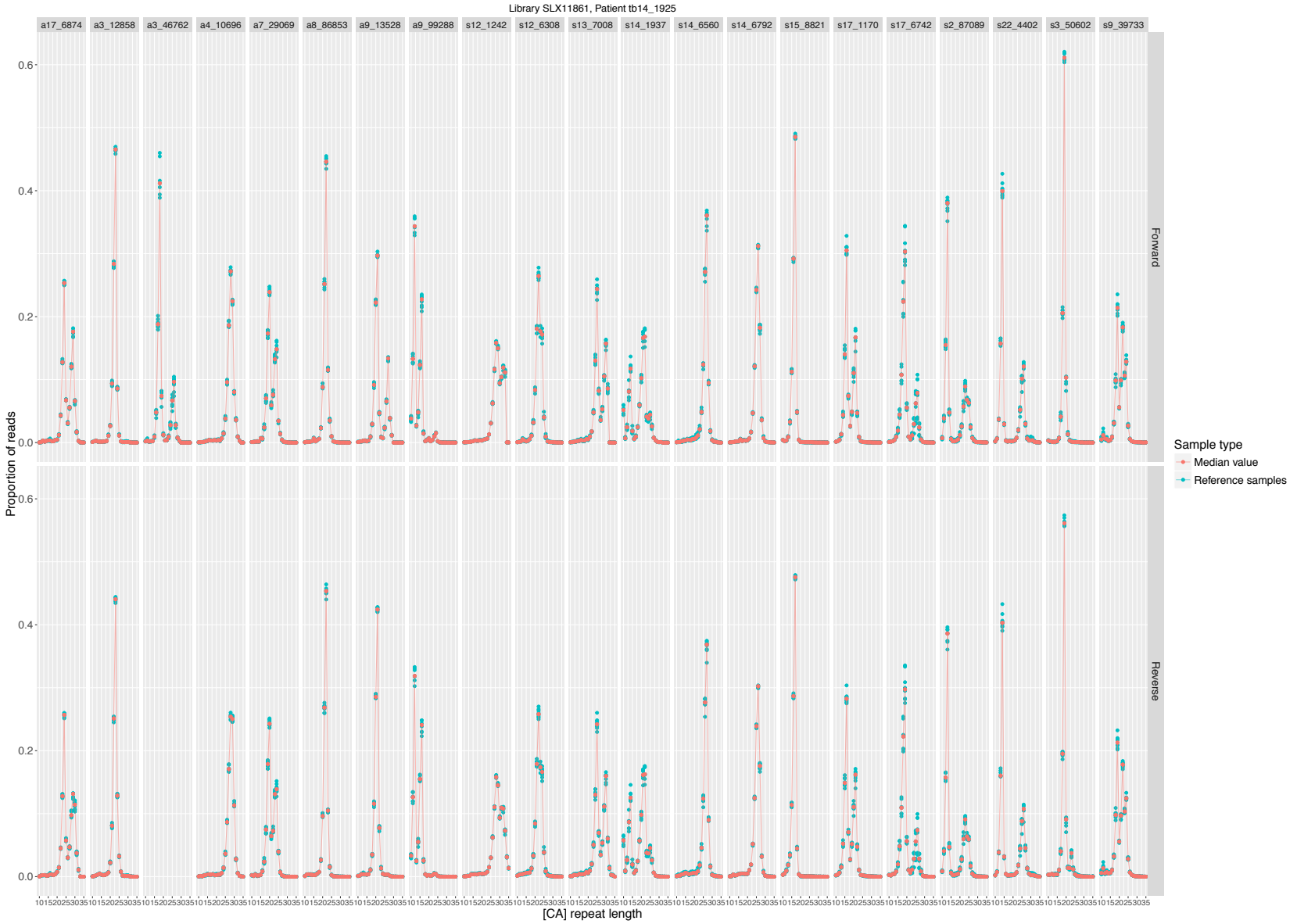

Fig. 6.3 Scatter plot showing reference distributions in one patient sample. 13 of the 21 loci display a bimodal distribution which is not amenable to mixture modelling analysis thus have to be filtered from downstream analysis. Locus a3_46762 has a missing reverse read distribution due to low read depth, as was also observed in the other patient.

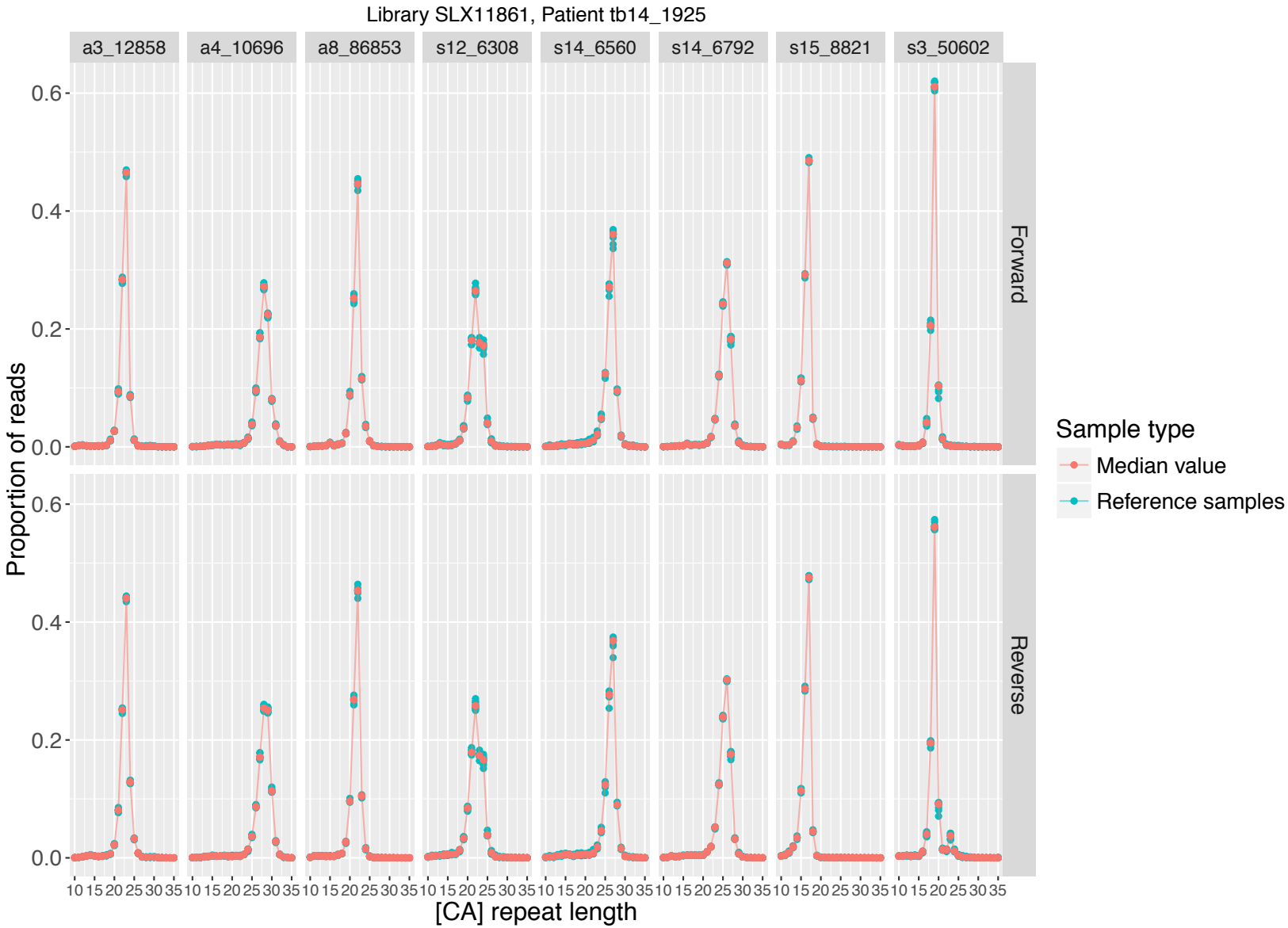

Fig. 6.4 Scatter plot showing reference distributions in patient sample after bimodal distributions were filtered out. Compared to mouse, the human samples have far more loci filtered due to the presence of heterozygous $[CA]_n$ alleles.

6.2.1 Setting Φ value thresholds for interpreting clone size in human crypts

The estimates of *shift* and Φ values were inferred as previously described in Chapter 4 using a mixture modelling analysis. Thresholds were defined on a locus by locus basis, as was done for mouse loci in Section 5.1.2. Firstly, crypts were categorised as either wild-type or mutant by setting a threshold at the tail end of the wild-type peak, Figure 6.5. This was possible for the majority of loci except for locus s9_39733 which lacked a distinct wild-type peak.

Two loci were homozygous in both patients and, therefore, amenable to analysis and comparison between the individuals. Out of these 2 loci, one of the loci (s15_8821) produced a clone size distribution as would be expected, Figure 6.6. The second locus, a8_86853, displayed an unusual disparity in Φ value distribution between the two patient samples. As these inferences were based on deviation from reference distributions specific to each patient, this disparity cannot be explained by a germline mutation and must instead have a somatic cause. In the 56 year old patient, virtually all of the crypts had a mutation at this locus whilst in the 74 year old patient, virtually all of the crypts contained a wild-type length microsatellite. It is possible that the younger patient had an increased microsatellite mutation rate in the tissue used for this study and this locus is a sentinel for this phenotype. Regardless, use of this locus leads to large differences in clone size estimates between the two patients and has to be considered with caution. For the purposes of this analysis and discussion, locus a8_86853 will be excluded.

Following the setting of a threshold for differentiating wild-type from mutant crypts, a second threshold was required to classify mutant crypts as PPC or WPC. In mouse, it was possible to use the increased incidence of WPCs in mismatch repair deficient crypts to define a threshold for differentiating between PPCs and WPCs. In human crypts, only locus a8_86853 had a discernible WPC peak but, due to the clone incidence disparity between patients, had to be filtered out. To set an approximate threshold, loci were categorised based on their wild-type peak distributions, Figure 6.7. Loci from wild-type mice were categorised in the same way, Figure 6.8. As accurate PPC-WPC thresholds had been ascertained for wild-type murine loci, the average threshold value for each category of murine loci was calculated (Category 1 $\Phi = 0.324$, Category 2 $\Phi = 0.327$ and Category 3 $\Phi = 0.315$). The human loci in the corresponding category had the PPC-WPC Φ value threshold set as the average value calculated in mouse. The final Φ value thresholds are summarised in Table

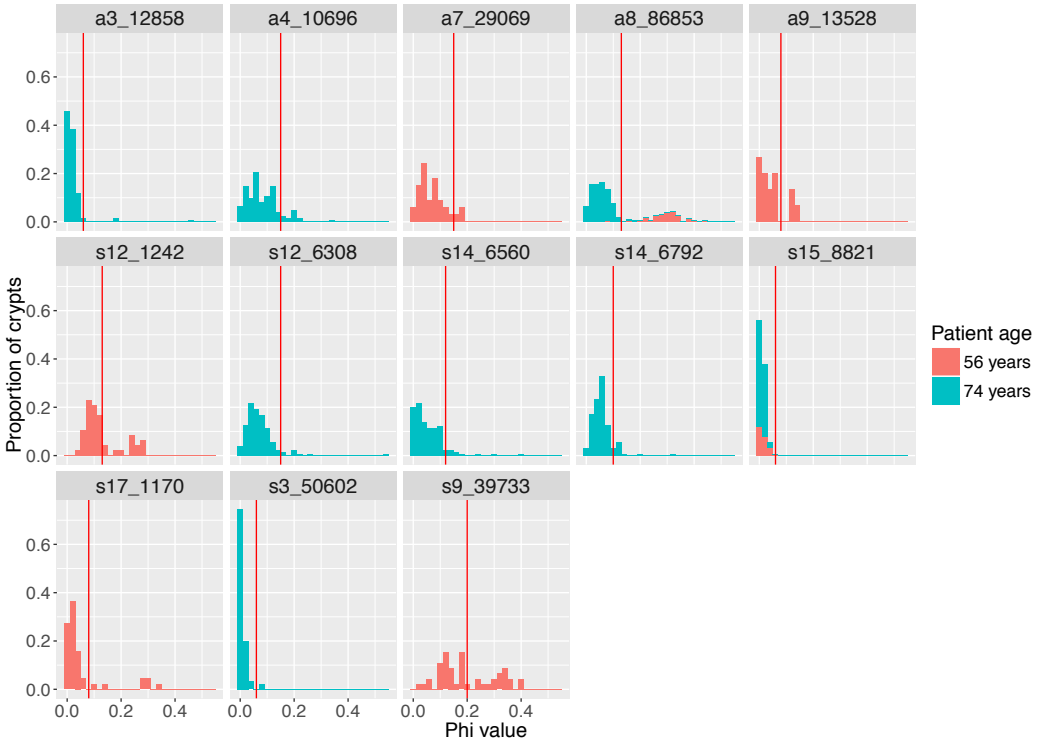

Fig. 6.5 Histogram showing frequency distribution of Φ values in human crypts from two patients. The red line shows the Φ value used to differentiate between wild-type and mutant crypts. Based on these plots, locus s9_39733 was excluded from future analyses.

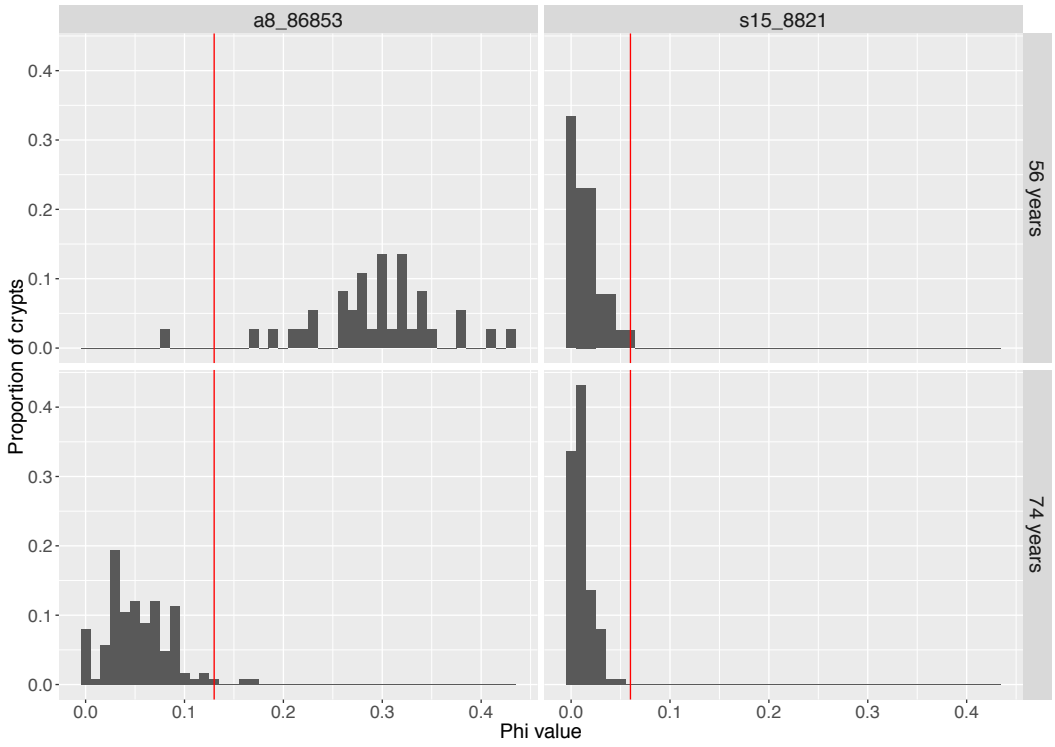

Fig. 6.6 Histogram showing frequency distribution of Φ values in human crypts at two loci studied in two patients. The red line shows the Φ value used to differentiate between wild-type and mutant crypts.There is a vast difference in the spread of Φ values between the two patients at a8_86853. Locus s15_8821 has a clone size distribution close to what would be expected. The difference in clone size distribution at locus a8_86853 is potentially due to a difference in mutation rate at that loci between the two patients.

6.2 and shown in Figure 6.9.

The number of crypts with a mutant locus was counted using the threshold defined for differentiating between wild-type and mutant crypts. As 4 loci were analysed in the 56 year old patient and 7 loci in the 74 year old patient, it is necessary to normalise the percentage mutant crypts to the number of loci analysed in each patient. The normalised percentage of mutant crypts is higher in the 56 year old patient (14.5% of crypts tested) compared with the 74 year old patient (5.9% of crypts tested), Figure 6.10. More mutated crypts would be expected in the older individual. Thus this result could be indicative of a mutator phenotype in the intestinal epithelium of the 56 year old patient though sampling error cannot be ruled out. Further crypt sequencing would be required to ascertain the normal level of clone incidence variation with patient age.

Given the age of the two patients, a much larger proportion of WPCs, relative to PPC frequency, would be expected in both patients. This may suggest that the use of the average murine threshold associated with wild-type peak distributions, for differentiating PPCs and WPCs, is not sufficient for correct interpretation of human data. For an accurate breakdown of PPCs and WPCs from human data to be performed, a larger dataset would be required to more accurately ascertain the different thresholds.

Despite the inability to accurately interpret the clone sizes, a significant proportion of crypts were shown to have mutant loci indicating that microsatellite sequencing can be used to quantify intra-cryptal clone size in human material. Observation of over two-fold more mutant crypts in the 56 year old patient in conjunction with the observation of increased mutation at locus a8_86853 is suggestive of an increased mutation rate in the epithelium of this patient. Further sequencing of crypts from this patient and consideration within the context of a larger study cohort would be required to shed light onto this hypothesis.

6.3 Discussion

In this chapter, I have described how the method developed for microsatellite sequencing from low template copy samples in mice can be translated to human material. The pipeline developed for the identification of microsatellites, of any nucleotide composition, and output of a file suitable for use in the online primer design tool, BatchPrimer, is readily applicable to any reference genome. BatchPrimer utilises multiple checks to predict primer efficacy such that the primers designed using this tool had a 79% and 91% success rate in mouse

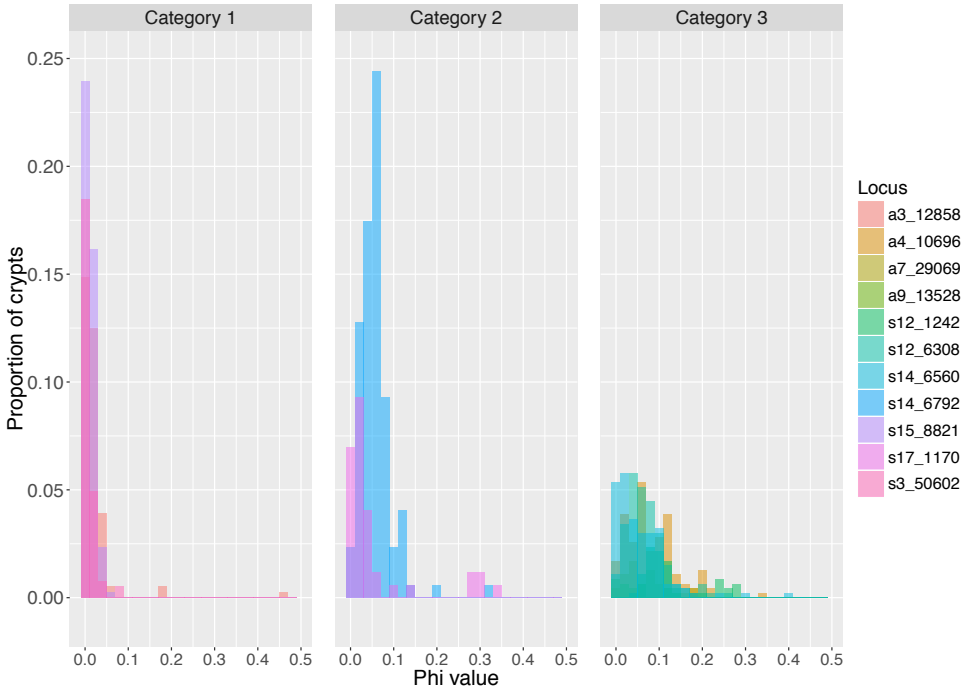

Fig. 6.7 Histogram showing frequency distribution of Φ values in human crypts. Each wild-type peak was used to group similar distributions together to form three categories. Equivalent distributions were sought in murine loci to enable approximation of a PPC-WPC threshold.

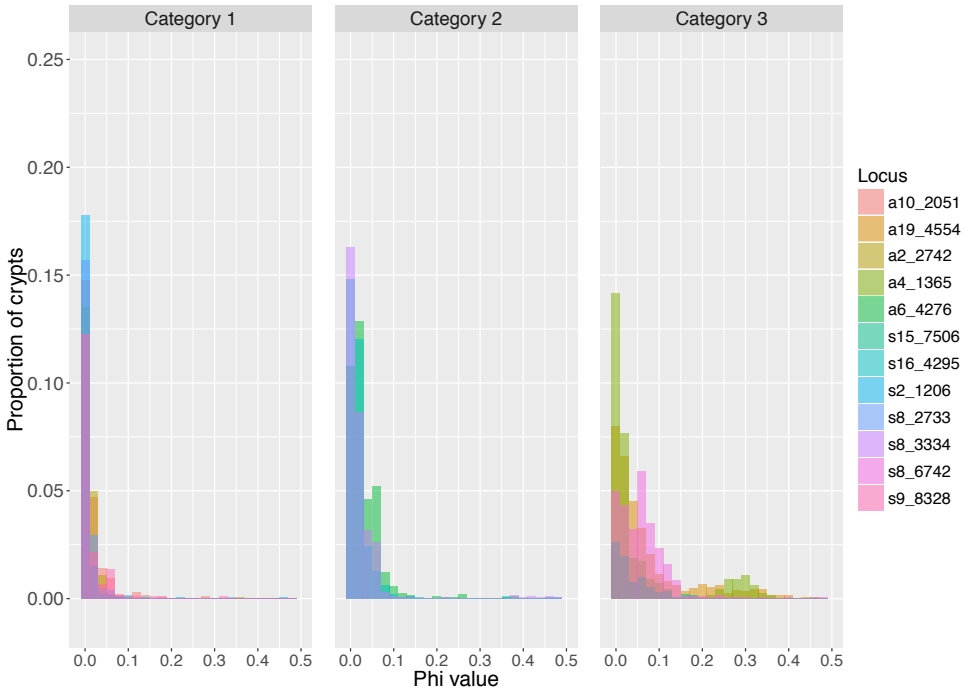

Fig. 6.8 Histogram showing frequency distribution of Φ values in wild-type mouse crypts. Each wild-type peak was used to group similar distributions together to form three categories in the same fashion as that done for human loci in Figure 6.7. The average PPC-WPC Φ threshold value for each category was calculated using the murine loci.

Locus	WT/PPC Threshold	PPC/WPC Threshold (Female)	PPC-WPC Threshold (Male)	Category
a3_12858	0.06	0.324	0.324	1
a4_10696	0.15	0.315	0.315	3
a7_29069	0.15	0.315	0.315	3
a9_13528	0.20	0.315	0.315	3
s12_1242	0.13	0.315	0.315	3
s12_6308	0.15	0.315	0.315	3
s14_6560	0.12	0.315	0.315	3
s14_6792	0.10	0.327	0.327	2
s15_8821	0.06	0.324	0.324	1
s17_1170	0.08	0.327	0.327	2
s3_50602	0.06	0.324	0.324	1

Table 6.2 Thresholds used to differentiate between wild-type, PPC and WPC crypts based on the inferred Φ value.

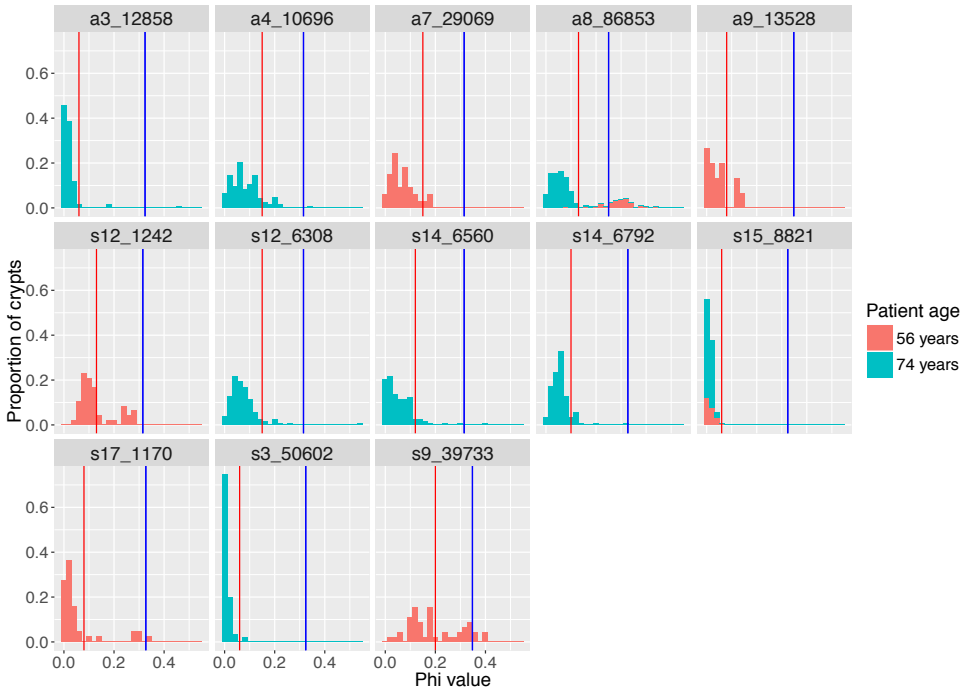

Fig. 6.9 Histogram showing frequency distribution of Φ values in human crypts from two patients. The red line shows the Φ value used to differentiate between wild-type and mutant crypts. The blue line shows the Φ value used to classify mutant crypts as PPC or WPC. Based on these plots, locus s9_39733 was excluded from future analyses and locus a8_86853 was excluded due to disparity in clone incidence between the two patients.

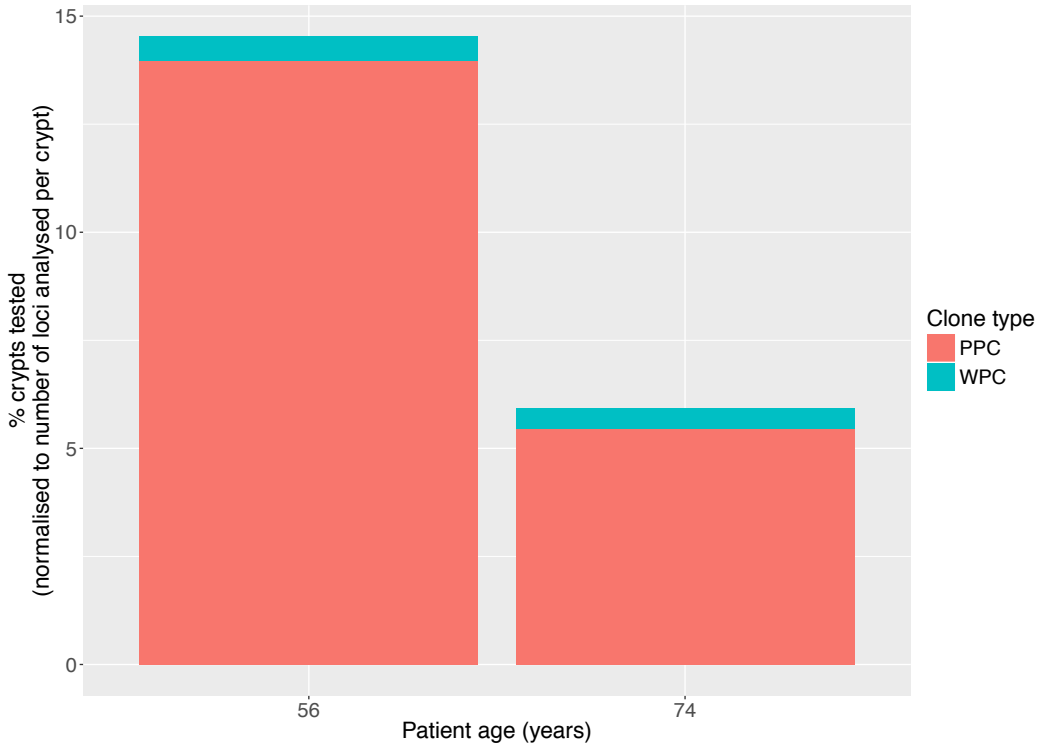

Fig. 6.10 Histogram showing percentage of crypts tested where a mutant loci was observed. Based on the thresholds set, including a PPC-WPC threshold set based on murine data, the predicted percentage of PPC and WPC are shown. A total of 48 and 128 crypts were sequenced in the 56 year old and 74 year old respectively. A total of 4 and 7 loci were analysed in the 56 year old and 74 year old respectively such that the percentage of crypts analysed was normalised to the number of loci analysed in each crypt.

and human respectively. It therefore seems more than likely that, if primers can be designed for these loci, they can be successfully amplified. This opens up the possibility of using this method in different species allowing for the multiplexed analysis of many individuals within many different species for purposes such as ecological and evolutionary research.

When translated from mouse to human material, the consistency in distribution between technical replicates was very high. This consistency allows for accurate reference distributions to be generated that can then be used for mixture modelling and estimation of Φ values for crypts of unknown clonal status. Being able to accurately infer Φ values, with appropriate interpretation of clonal status, in many crypts from humans at different ages will allow for observation of the age related change in clone size distribution in human tissues. Using the same model described in the Kozar et al [54] study, the functional stem cell number and stem cell replacement rate within the human colonic crypt could be inferred.

As well as its applicability to quantifying stem cell dynamics, the consistency between replicates allows for comparison of genotypes between two samples of unknown relatedness. This could be applied for straightforward genotyping requirements such as parental testing. However, the ability to also be able to detect and quantify the presence of minor microsatellite species at differing lengths could allow applicability to forensic samples. Overall, the ability to use Next Generation Sequencing technologies to perform these sorts of analyses from low template samples could be an incredibly useful tool for forensic studies and genotyping applications.

From sequencing of reference material from patient samples, a disparity between reference genome microsatellite length calls and the length calls seen in patient samples was uncovered. The source of this discrepancy is likely to be due to sequencing methods used to generate reference genomes which have poor fidelity when sequencing repetitive regions of the genome. The method described in this dissertation provides a high fidelity method for calling of microsatellite length at multiple loci in many different individuals. It could therefore be possible to generate more accurate microsatellite sequence information for reference genomes using this method.

When sequencing microsatellites in mice, the majority of the microsatellites were homozygous in length and little length variation was seen between individuals. In contrast, the human sequencing data revealed large polymorphism between individuals and heterozygosity of microsatellite length at multiple loci. As a result, many of the loci sequenced in human samples had to be removed from the analysis. This is an inefficient use of these

datasets. Instead, the computational pipeline used to analyse this data could be adapted to deal with loci with bimodal distributions. Alternatively, the multiplex groups used in future studies could be expanded further so that redundancy in analysable loci does not greatly affect the overall power of the method. Ultimately, a combination of improved computational methods and expanded multiplex groups would be the ideal solution for dealing with the prevalence of heterozygosity in human samples.

Interestingly, one locus showed a large disparity in Φ value distribution between the two patients studied. Understanding the cause of this disparity would require further study but could potentially be indicating an increased microsatellite mutation rate in the normal epithelium of this patient with locus a8_86853 acting as a sentinel locus. This could also explain the increased number of mutant crypts observed in this patient. Larger cohorts and, preferably, sequencing of more material from this individual will enable greater insight into this hypothesis.

The study presented here serves as a proof of principle for the quantification of clone size within human intestinal crypts. In order to infer accurate stem cell dynamic metrics, larger patient cohorts would be required. The use of the HiSeq 4000 and a custom sequencing adaptor allows for expansion of the multiplexing capabilities of the approach which will significantly lower the cost of any future study using this method. Overall, this method provides an exciting opportunity to use an unbiased, transgene-free method for the study of intestinal stem cell dynamics in humans.

Chapter 7

Discussion

Since the early pioneers of modern cellular biology in the late 19^{th} century, much has been learned and described of the homeostatic turnover of cells in all tissues of the human body. The laboratory mouse has played a central role in these studies to allow unprecedented insights into the workings of healthy and diseased tissues. The intestinal epithelium, with its notable mitotic activity, has been a model system for the study of epithelium and cell kinetics. In recent years, there has been a focus on the discovery of molecular markers for different cell populations that has been fruitful in aiding the observation and description of the underlying cellular behaviour of the crypt of Leiburkühn.

However, restriction of study to marked populations can lack specificity and can lead to an unhelpful assumption that a stem cell marker associating with a particular cell state is required for effective stem cell research. The development of the Rosa26-$[CA]_{30}$-eYFP and Rosa26-$[CA]_{30}$-SynBglA mice, provide a system independent of such markers and chemical induction for the functional study of intestinal stem biology in mouse [54]. Inference of functional stem cell number and stem cell replacement rate in these mice, revealed notably fewer functional stem cells than previously thought and regional differences in stem cell dynamics along the murine intestinal tract. Recent advances in DNA sequencing technologies presented an opportunity to use sequencing of endogenous $[CA]_{30}$ tracts to perform the same analysis in human tissues.

Existing strategies for studying human intestinal stem cell dynamics largely rely upon the loss of a protein that can be detected and tracked using immunohistochemical, and related, techniques [17, 96]. Though these approaches are useful in the study of clonal patches within the intestinal epithelium [40, 50], the neutrality of the loss of such proteins is currently unknown. Thus inferences about clonal drift and stem cell dynamics may be biased

by the loss of such markers. The use of microsatellite length change provides a neutral marker for clones within the intestinal epithelium.

In this dissertation, I have described multiple experiments validating the use of microsatellite sequencing for clone size quantification in murine and human crypts. Use of synthetic loci and murine tissues allowed for validation of the method in well described systems. In addition, an analysis pipeline was developed that utilises Illumina sequencing data from single crypts to accurately infer clone size. Furthermore, through study of Msh2 deficient epithelium, microsatellite mutation rates at endogenous loci were calculated and the spectrum of mutations, supporting a loop insertion-deletion mechanism of microsatellite mutation, were observed. Finally, I provide evidence for the translation of this technique into human tissues and the quantification of clone size in patient material.

The development of a multiplexed microsatellite sequencing protocol for single crypts posed many technical challenges. Initially, a protocol was required for the reliable and consistent isolation of single crypts. By using low adherence equipment and dispensing crypts directly into lysis buffer, the technique for single crypt isolation was significantly improved. Perhaps unsurprisingly, the use of paraformaldehyde fixation prior to crypt isolation rendered the DNA unsuitable for amplification. Development of a reliable technique for single crypt isolation has the potential for use in a range of experiments relating to single crypt analyses.

Furthermore, steps were required to reduce the level of DNA contamination in single crypt lysates. The potential sources of DNA contamination are: 1) cell free DNA contained within the single crypt suspension media, 2) cell debris contained within the single crypt suspension media and 3) constitutive contamination from surrounding cells that do not contribute to epithelial lineages within the crypt. All of these sources are likely to lead to contribution of microsatellite loci at wild type lengths leading to under estimation of mutant clone size. Quantification of cell free DNA allowed for selection of optimal single crypt suspension conditions. The passage of each single crypt through PBS droplets, so called 'crypt washing', allowed for visual inspection for cell debris and selection of crypts with minimal stromal attachment. Additionally, crypt washing was shown in later experiments to be a powerful tool for the reduction of DNA contamination in crypt lysate and significant improvement in mutant clone detection. These experiments were critical in improving the sensitivity and reliability of the method. Again, these observations lead to critical improvements in any experiment relating to single crypt analysis.

Loop insertion-deletion is proposed to be the mechanism of microsatellite mutation *in vivo*. A similar process of PCR 'stuttering' has been described for regions containing microsatellites suggesting that the same mutational process is present *in vitro*. The presence of *in vitro* polymerase slippage leading to reduced fidelity in the amplification of microsatellites is a critical issue for the accurate inference of clone size from single crypt lysates. The key breakthrough in addressing this issue was the observation of high fidelity and high efficacy of the New Engand Biolabs Phusion DNA polymerase in amplifying microsatellites from mouse genomic DNA. Intriguingly, the DNA polymerases that performed best at microsatellite amplification contained the Sso7d domain commonly added to DNA polymerases to improve processivity. It is possible that the increased contact area between the enzyme and DNA improves fidelity in highly repetitive stretches of DNA but this remains to be formally addressed.

In addition to error during PCR, prior studies on sequencing of microsatellites using Illumina technologies focussed on shorter microsatellites or those containing tetranucleotide repeats that are commonly used for human genotyping studies. These shorter and more complex microsatellites have lower mutation rates and are less affected by loop insertion-deletion mutagenesis. Therefore, prior to this study, the performance of the Illumina sequencers in sequencing dinucleotide microsatellites was unknown. As can be seen from the comparison of MiSeq and HiSeq sequencing data, there is a slightly different error profile indicative of sequencing error contribution to overall error in microsatellite length calling. The exact contribution of this error to the overall error remains to be quantified but, given the relatively small amount of amplification present in Illumina sequencing relative to PCR amplification, it seems likely that the majority of the error occurs during PCR with only a small contribution from sequencing.

Using the HiSeq leads to a substantial increase in average read depth per amplicon compared to the MiSeq which results in improved replicate consistency. Accurately calling mutant distributions hinges upon the ability to reproducibly capture a reference distribution with a high level of accuracy. With high levels of consistency between wild-type replicates, it is possible to more confidently call deviations from this distribution as mutant and therefore biologically meaningful. The high read depths used in this study minimises the variability seen between replicates so that the only observable source of variation is that caused by technical noise and/or biological differences in microsatellite length. Overall, this will improve the sensitivity of the assay and accuracy of clone size estimates.

The development of plasmids containing murine genomic regions with differing lengths of $[CA]_n$ microsatellites were a valuable tool for *in vitro* validation of the microsatellite sequencing protocol. Initially, qPCR analysis was done to show that there was no amplification bias towards microsatellites at shorter lengths. Mixing experiments show the ability of microsatellite sequencing to deconvolute wild-type and mutant mixtures. Additionally, mixes, particularly 50:50 mixes, of wild-type and mutant microsatellites displayed no amplification bias suggesting that small changes in microsatellite length do not lead to amplification bias even at low template copies. Interestingly, estimates of mutant read proportion, Φ, display a marked increase in variability when copies are reduced from 500 to 165 plasmid copies suggesting that allele dropout contributes significantly to Φ estimate error at low template copies. This would be a major consideration when considering adapting this protocol to samples with lower template copies, such as single cell experiments.

Current methods for analysing microsatellite sequencing data are tailored to estimating the dominant microsatellite length such that any information about minor microsatellite lengths are assumed to be artifactual and discarded. As this minor microsatellite length is crucial for inferring clone size, a custom script was designed to measure the length of the [CA] repeat in each read contained within the sequencing data and compile a matrix displaying the number of reads containing [CA] repeats of differing lengths. Output from this [CA] counting tool could then be used for analysis of mutant microsatellite length contribution, in a locus and crypt specific manner. Mixture modelling was used to infer this contribution. An assumption of the mixture model was that a shifted reference distribution would accurately predict the mutant distribution. This was shown not be the case at some loci leading to inaccuracies in Φ estimation. To account for this, the threshold for differentiating between wild-type, partly populated and wholly populated crypts had to be adjusted on a locus by locus basis.

Key to determining these thresholds was the use of Φ estimate distributions from wild-type and Msh2 deficient crypts. A threshold for differentiating between wild-type and mutant crypts can be achieved by setting a threshold at the tail end of the wild-type crypt distribution. A second threshold is required to classify mutant crypts as either PPC or WPC. In Msh2 deficient crypts, at 70 days post-induction of Msh2 knockout, a peak representing wholly populated crypts is clearly observable. Through visually inspecting each locus, a second threshold could be set at the tail end of the WPC peaks. These thresholds can then be used to call wild-type, PPCs and WPCs in wild-type epithelium and at a range of

time points post-induction of Msh2 knockout. Though this is a viable option for murine studies, collection of mismatch repair deficient human crypts is unfeasible such that larger sequencing cohorts will be required before accurate PPC-WPC thresholds can be set.

In mouse, it is possible to use simulated data to predict the clone size distribution expected at different microsatellite mutation rates. The observation of WPC accumulation in Msh2 deficient crypts robustly matches that predicted from simulation of clone size distribution with a microsatellite mutation rate 175-fold higher than that seen in wild-type epithelium. The observed incidence of PPCs was slightly lower than predicted from the simulated data. This underestimate of PPCs can be explained by the relative insensitivity of the method in detecting clones at 2/7 size or smaller. Adjustment of the model to account for this phenomena lead to close matching of the observed and predicted distributions. This highlights robust detection of clones approximately 3/7 in size and larger with some difficulties detecting smaller clones.

As more data is accrued, distributions observed in WPCs could be used as mutant distributions. This would remove the need for prediction of mutant distributions, using wild-type reference distribution shifts, and reduce the error observed in Φ estimate. This would allow for thresholds to be set closer to the extremes of possible Φ values and increase the dynamic range of the assay which ultimately improves accuracy of clone size estimates. That being said, the current method appears to be highly effective in differentiating between wild-type, partly populated and wholly populated crypts such that the methods described above would be subtle improvements rather than required modifications.

The current protocol is well positioned to be scaled. Firstly, the use of a custom sequencing adaptor set allows for expansion of the index set beyond the current limit of 384 unique indexes by simply synthesising a new set of adaptors with an increased number of unique indexes. Secondly, the amenability of this protocol to HiSeq sequencing allows for large amounts of data to be accrued in relatively short time scales. The only limit to this would be ensuring all loci in all crypts are above the 1000 reads limit before signal to noise becomes an issue. The current bottleneck in the scaling of this protocol lies in the collection of single crypts. Currently, all crypts have to be manually picked and have an upper limit of 200 crypts per day per person. Adaptation of large particle flow cytometry protocols, such as those currently used for sorting *C. elegans*, or the design of a custom microfluidic device would be invaluable in the scaling of this protocol. Nonetheless, the current protocol has a relatively high throughput allowing for hundreds of crypts to be isolated, lysed and prepared

for sequencing in 3-4 days.

Based on the observations of the transgenic $[CA]_{30}$ microsatellite in the Rosa26-$[CA]_{30}$-synBglA mouse, the frequency of PPCs and WPCs observed at endogenous $[CA]_{30}$ microsatellites was lower than predicted. The likely cause of this is a reduced mutation rate at these loci. As the transgenic $[CA]_{30}$ is under a housekeeping promoter (*Rosa26*), it is possible that transcriptional activity may influence microsatellite mutation rate. In Msh2 deficient epithelium, the transgenic $[CA]_{30}$ locus is ranked as one of the most mutable loci, consistent with our prediction that this locus displays an increased mutation rate. This is particularly notable when considering that this locus is monoallelic such that the mutation rate is half that expected at a biallelic locus. All but one of the other loci studied in mouse were biallelic and the only other monoallelic locus had the lowest inferred mutation rate.

In addition to studying intra-cryptal clone size variation in wild-type epithelium, clone size distributions were quantified in Msh2 deficient epithelium also. The VillinCreERT; $Msh2^{fl/fl}$ mouse was utilised to induce Msh2 deficiency using tamoxifen. From this analysis, expected clone size distributions were observed and the average mutation rate for endogenous microsatellites in Msh2 deficiency was calculated to be $1.93x10^{-2}$ per mitosis per locus or $9.63x10^{-3}$ mutation per mitosis per allele. This calculation shows the possibility of using this method to infer mutation rates at many different loci based on known values for functional stem cell number and stem cell replacement rate in the murine intestinal epithelium. This could potentially be extended to other applications such as inferring the rate of single nucleotide and structural variations. Though different approaches to amplifying DNA from single crypt lysate would almost certainly be required.

During initial sequencing of loci from mouse genomic DNA, very stable microsatellite lengths were observed across the majority of loci. Somewhat unexpectedly, five of the loci were variable in length between individuals, termed 'germline variable' loci. We hypothesised that these loci may be more unstable and, therefore, have a higher mutation rate somatically. Primer pairs for these loci were included in the final multiplex group and analysed in both wild-type and Msh2 deficient crypts. Though the mutation rate in wild-type epithelium was too low to comment meaningfully on these loci, these loci had an elevated mutation rate in Msh2 deficient epithelium. So much so that defining thresholds for wild-type, PPC and WPC was unfeasible at the majority of these loci. This indicates that these loci could be highly informative for quantifying clone size in wild-type epithelium but would be unfit for this purpose in a DNA repair deficient setting such as in tumours or patients with germline

mutations in such pathways.

The use of germline variability as a means of predicting somatically mutable loci is a useful tool in mice as, due to significant levels of inbreeding, these loci tend to be homozygous in length despite increased germline variability. However, if such germline variable loci exist in human populations, they are highly likely to be heterozygous and would be unsuitable for the current analysis pipeline. Further development of the analysis pipeline to allow use of heterozygous loci would enable testing of somatic mutability in germline variable loci in humans.

The mechanism of microsatellite mutation is thought to be through loop insertion-deletion leading to small scale, iterative changes in microsatellite length [46, 69, 107]. Observation of the mutational shifts in wild-type epithelium supports this mechanism with a significant bias towards smaller scale mutations. Interestingly, the bias of shifts differs between loci and it seems likely the genomic context of the microsatellite influences the pattern of mutations observed. Furthermore, observation of the same loci in Msh2 deficient crypts revealed the same pattern of small scale changes. This supports the notion that there is dynamic interplay between mutation and repair at microsatellites thus when repair mechanisms are deficient there is a bias towards mutation. The nature of this mutation appears to genomically encoded and is independent of any other mutational process in Msh2 deficient epithelium.

Following *in vitro* and *in vivo* validation of microsatellite sequencing for clone size quantification in mouse, the feasibility of translation to human tissue was assessed. The increased DNA content of single human crypts theoretically aided the translation of this protocol to human crypts. Indeed, it is possible that the number of PCR cycles could be reduced to improve amplification fidelity of human crypt lysate in future. Furthermore, initially optimising the protocol in mouse informed many of the steps in optimising the protocol for human crypts such as primer design and optimal primer concentrations. Thus, translation of this protocol from murine material to human material was successful.

These observations pave the way for further investigations of stem cell dynamics in human colonic material. It should also be possible to quantify stem cell dynamics in diseased colon also. This could allow for the quantification of stem cell dynamics in patients suffering from inflammatory bowel disease or in a dysplastic adenoma. Insights into the alteration of stem cell dynamics in the diseased setting could shed light onto: 1) the role of perturbed stem cell dynamics in the aetiology and pathogenesis of disease; 2) identification of individuals at heightened risk of colorectal cancer; 3) the potential for targeting perturbed

cellular kinetics to specifically remove a pathological cell population or develop effective chemopreventive therapies.

This method also has the potential for wider application beyond quantifying stem cell dynamics. Firstly, the ability to perform massively multiplexed genotyping could allow for adoption in fields such as forensic science or in parental testing [74]. Secondly, the ability to infer kinship based on microsatellite length change could be used to inform hierarchical clustering at a population level in evolutionary or ecological studies [43]. Finally, the identification of loci with higher mutation rates suggests that this method could be used to ascertain hypermutability in sentinel loci. Sentinel loci are already used to ascertain the presence of microsatellite instability in colorectal cancers [8, 56]; the method described in this dissertation could allow for testing with higher sensitivity and without the requirement for large template copies, such as from circulating tumour DNA or fecal samples. This could potentially be used to identify patients with loss of mismatch repair in the colon through existing screening programmes.

In addition to sequencing microsatellites, the multiplexed protocol could potentially be used to sequence known cancer mutation hotspots in genes such as KRAS or BRAF [19, 25]. With known parameters for functional stem cell number and stem cell replacement rate, it would be possible to infer the mutation rates and spectrum of mutation at these loci. Furthermore, it is plausible that if a large enough cohort was studied, it would be possible to infer any biases in the maintenance of these mutations in normal epithelium, indicative of biased drift, a phenomena already described in murine intestine.

Overall, evidence for the potential of this novel technique in the quantification of stem cell dynamics in an unbiased, transgene-free way has been presented. This method is amenable to use in large scale sequencing projects to ascertain human colonic stem cell dynamics. Furthermore, the isolation, sequencing and analysis pipeline has the flexibility such that each step has the potential for scalability to reduce the time and financial investment required for the application of this method to a wide range of applications beyond that in quantifying intestinal stem cell dynamics.

References

[1] A. Adey, H. G. Morrison, A. Asan, X. Xun, J. O. Kitzman, E. H. Turner, B. Stackhouse, A. P. MacKenzie, N. C. Caruccio, X. Zhang, and J. Shendure. Rapid, low-input, low-bias construction of shotgun fragment libraries by high-density in vitro transposition. *Genome Biology*, 11(12):R119, 2010.

[2] A.-M. Baker, B. Cereser, S. Melton, A. G. Fletcher, M. Rodriguez-Justo, P. J. Tadrous, A. Humphries, G. Elia, S. A. McDonald, N. A. Wright, B. D. Simons, M. Jansen, and T. A. Graham. Quantification of Crypt and Stem Cell Evolution in the Normal and Neoplastic Human Colon. *Cell Reports*, 8:940–947, 2014.

[3] B. A. Baptiste, G. Ananda, N. Strubczewski, A. Lutzkanin, S. J. Khoo, A. Srikanth, N. Kim, K. D. Makova, M. M. Krasilnikova, and K. A. Eckert. Mature microsatellites: mechanisms underlying dinucleotide microsatellite mutational biases in human cells. *G3*, 3:451–463, 2013.

[4] N. Barker, J. H. van Es, J. Kuipers, P. Kujala, M. van den Born, M. Cozijnsen, A. Haegebarth, J. Korving, H. Begthel, P. J. Peters, and H. Clevers. Identification of stem cells in small intestine and colon by marker gene Lgr5. *Nature*, 449(7165):1003–7, 2007.

[5] A. J. Becker, E. A. McCulloch, and J. E. Till. Cytological Demonstration of the Clonal Nature of Spleen Colonies Derived from Transplanted Mouse Marrow Cells. *Nature*, 197(4866):452–454, 1963.

[6] D. Bentley, S. Balasubramanian, H. Swerdlow, G. Smith, J. Milton, C. Brown, K. Hall, D. Evers, C. Barnes, H. Bignell, J. Boutell, J. Bryant, R. Carter, R. Keira Cheetham, A. Cox, D. Ellis, M. Flatbush, N. Gormley, S. Humphray, L. Irving, M. Karbelashvili, S. Kirk, H. Li, X. Liu, K. Maisinger, L. Murray, B. Obradovic, T. Ost, M. Parkinson, M. Pratt, I. M. Rasolonjatovo, M. Reed, R. Rigatti, C. Rodighiero, M. Ross, A. Sabot, S. Sankar, A. Scally, G. Schroth, M. Smith, V. Smith, A. Spiridou, P. Torrance, S. Tzonev, E. Vermaas, K. Walter, X. Wu, L. Zhang, M. Alam, C. Anastasi, I. Aniebo, D. M. Bailey, I. Bancarz, S. Banerjee, S. Barbour, P. Baybayan, V. Benoit, K. Benson, C. Bevis, P. Black, A. Boodhun, J. Brennan, J. Bridgham, R. Brown, A. Brown, D. Buermann, A. Bundu, J. Burrows, N. P. Carter, N. Castillo, M. Chiara E Catenazzi, S. Chang, R. Neil Cooley, N. Crake, O. Dada, K. Diakoumakos, B. Dominguez-Fernandez, D. Earnshaw, U. Egbujor, D. Elmore, S. Etchin, M. Ewan, M. Fedurco, L. Fraser, K. Fuentes Fajardo, W. Scott Furey, D. George, K. Gietzen, C. Goddard, G. Golda, P. Granieri, D. Green, D. Gustafson, N. Hansen, K. Harnish, C. Haudenschild, N. Heyer, M. Hims, J. Ho, A. Horgan, K. Hoschler, S. Hurwitz, D. Ivanov, M. Johnson, T. James, T. A. Huw

Jones, G. Kang, T. Kerelska, A. Kersey, I. Khrebtukova, A. Kindwall, Z. Kingsbury, P. Kokko-Gonzales, A. Kumar, M. Laurent, C. Lawley, S. Lee, X. Lee, A. Liao, J. Loch, M. Lok, S. Luo, R. Mammen, J. Martin, P. Mccauley, P. Mcnitt, P. Mehta, K. Moon, J. Mullens, T. Newington, Z. Ning, B. Ling Ng, S. Novo, M. J. O'Neill, M. Osborne, A. Osnowski, O. Ostadan, L. Paraschos, L. Pickering, A. Pike, D. Chris Pinkard, D. Pliskin, J. Podhasky, V. Quijano, C. Raczy, V. Rae, S. Rawlings, A. Chiva Rodriguez, P. Roe, J. Rogers, M. Rogert Bacigalupo, N. Romanov, A. Romieu, R. Roth, N. Rourke, S. Ruediger, E. Rusman, R. Sanches-Kuiper, M. Schenker, J. Seoane, R. Shaw, M. Shiver, S. Short, N. Sizto, J. Sluis, J. Ernest Sohna Sohna, E. Spence, K. Stevens, N. Sutton, L. Szajkowski, C. Tregidgo, G. Turcatti, S. Vandevondele, Y. Verhovsky, S. Virk, S. Wakelin, G. Walcott, J. Wang, G. Worsley, J. Yan, L. Yau, M. Zuerlein, J. Mullikin, M. E. Hurles, N. Mccooke, J. West, F. Oaks, P. Lundberg, D. Klenerman, R. Durbin, and A. Smith. Accurate whole human genome sequencing using reversible terminator chemistry. *Nature*, 456(7218):53–59, 2008.

[7] M. Bjerknes and H. Cheng. Methods for the isolation of intact epithelium from the mouse intestine. *The Anatomical record*, 199(4):565–574, 1981.

[8] C. R. Boland and A. Goel. Microsatellite Instability in Colorectal Cancer. *Gastroenterology*, 138(6):2073–2087, 2010.

[9] V. Bonadona, B. Bonaiti, S. Olschwang, S. Grandjouan, L. Huiart, M. Longy, R. Guimbaud, B. Buecher, Y.-J. Bignon, O. Caron, C. Colas, C. Nogues, S. Lejeune-Dumoulin, L. Olivier-Faivre, F. Polycarpe-Osaer, T. D. Nguyen, F. Desseigne, J.-C. Saurin, P. Berthet, D. Leroux, S. Manouvrier, T. Frebourg, H. Sobol, C. Lasset, C. Bonaiti-Pellie, and F. C. G. Network. Cancer Risks Associated With Germline Mutations in MLH1, MSH2, and MSH6 Genes in Lynch Syndrome. *JAMA*, 305(22):2304–10, 2013.

[10] D. M. Bornman, M. E. Hester, J. M. Schuetter, M. D. Kasoji, A. Minard-Smith, C. A. Barden, S. C. Nelson, G. D. Godbold, C. H. Baker, B. Yang, J. E. Walther, I. E. Tornes, P. S. Yan, B. Rodriguez, R. Bundschuh, M. L. Dickens, B. A. Young, and S. A. Faith. Short-read, high-throughput sequencing technology for STR genotyping. *BioTechniques Rapid Dispatches*, 1:1–6, 2012.

[11] D. T. Breault, I. M. Min, D. L. Carlone, L. G. Farilla, D. M. Ambruzs, D. E. Henderson, S. Algra, R. K. Montgomery, A. J. Wagers, and N. Hole. Generation of mTert-GFP mice as a model to identify and study tissue progenitor cells. *Proc Nat Acad Sci USA*, 105(30):10420–5, 2008.

[12] S. J. A. Buczacki, H. I. Zecchini, A. M. Nicholson, R. Russell, L. Vermeulen, R. Kemp, and D. J. Winton. Intestinal label-retaining cells are secretory precursors expressing Lgr5. *Nature*, 495(7439):65–9, 2013.

[13] B. Budowle, A. J. Eisenberg, and A. van Daal. Validity of low copy number typing and applications to forensic science. *Croatian medical journal*, 50(3):207–217, 2009.

[14] J. M. Butler. Short tandem repeat typing technologies used in human identity testing. *BioTechniques*, 43(4), 2007.

[15] J. M. Butler, E. Buel, F. Crivellente, and B. R. McCord. Forensic DNA typing by capillary electrophoresis using the ABI Prism 310 and 3100 genetic analyzers for STR analysis. *Electrophoresis*, 25(10-11):1397–412, 2004.

[16] J. Cairns. Mutation selection and the natural history of cancer. *Nature*, 255:197–200, 1975.

[17] F. Campbell, G. T. Williams, M. A. Appleton, M. F. Dixon, M. Harris, and E. D. Williams. Post-irradiation somatic mutation and clonal stabilisation time in the human colon. *Gut*, 39(4):569–573, 1996.

[18] K. D. Carlson, P. H. Sudmant, M. O. Press, E. E. Eichler, J. Shendure, and C. Queitsch. MIPSTR: A method for multiplex genotyping of germline and somatic STR variation across many individuals. *Genome Research*, 125(5):750–761, 2015.

[19] Y. S. Chang, I. L. Lin, K. T. Yeh, and J. G. Chang. Rapid detection of K-, N-, H-RAS, and BRAF hotspot mutations in thyroid cancer using the multiplex primer extension. *Clinical Biochemistry*, 46(15):1572–1577, 2013.

[20] H. Cheng. Origin, differentiation and renewal of the four main epithelial cell types in the mouse small intestine I. Columnar cells. *Am J Anat*, 141:461–480, 1974.

[21] H. Cheng. Origin, Differentiation and Renewal of the Four Main Epithelial Cell Types in the Mouse Small Intestine II. Mucous Cells. *Am J Anat*, 141:481–502, 1974.

[22] H. Cheng. Origin, differentiation and renewal of the four main epithelial cell types in the mouse small intestine IV. Paneth Cells. *Am J Anat*, 141:521–536, 1974.

[23] H. Cheng and C. P. Leblond. Origin, Differentiation and Renewal of the Four Main Epithelial Cell Types in the Mouse Small Intestine III. Entero-endocrine cells. *Am J Anat*, 141:503–520, 1974.

[24] H. Cheng and C. P. Leblond. Origin, differentiation and renewal of the four main epithelial cell types in the mouse small intestine V. Unitarian Theory Of The Origin Of The Four Epithelial Cell Types. *Am J Anat*, 141:537–562, 1974.

[25] B. N. Cyriac Kandoth, Michael D. McLellan, Fabio Vandin, Kai Ye and C. Lu. Mutational landscape and significance across 12 major cancer types. *Nature*, 503(7471):333–339, 2013.

[26] F. Diaz. Cytochrome c oxidase deficiency: patients and animal models. *Biochimica et biophysica acta*, 1802(1):100–110, 2010.

[27] F. El Marjou, K. P. Janssen, B. H. J. Chang, M. Li, V. Hindie, L. Chan, D. Louvard, P. Chambon, D. Metzger, and S. Robine. Tissue-specific and inducible Cre-mediated recombination in the gut epithelium. *Genesis*, 39(3):186–193, 2004.

[28] H. Ellegren. Microsatellites: simple sequences with complex evolution. *Nat Rev Genet*, 5(6):435–445, 2004.

[29] M. Escobar, P. Nicolas, F. Sangar, S. Laurent-Chabalier, P. Clair, D. Joubert, P. Jay, and C. Legraverend. Intestinal epithelial stem cells do not protect their genome by asymmetric chromosome segregation. *Nature communications*, 2(258):1–9, 2011.

[30] H. F. Farin, I. Jordens, M. H. Mosa, O. Basak, J. Korving, D. V. F. Tauriello, K. de Punder, S. Angers, P. J. Peters, M. M. Maurice, and H. Clevers. Visualization of a short-range Wnt gradient in the intestinal stem-cell niche. *Nature*, 530(7590):340–343, 2016.

[31] A. Fazekas, R. Steeves, and S. Newmaster. Improving sequencing quality from PCR products containing long mononucleotide repeats. *BioTechniques*, 48(4):277–85, 2010.

[32] L. C. Fernández, M. Torres, and F. X. Real. Somatic mosaicism: on the road to cancer. *Nature Reviews Cancer*, 16(1):43–55, 2015.

[33] T. Fevr, S. Robine, D. Louvard, and J. Huelsken. Wnt / β -Catenin Is Essential for Intestinal Homeostasis and Maintenance of Intestinal Stem Cells. *Mol Cell Biol*, 27(21):7551–7559, 2007.

[34] J. W. Fondon, A. Martin, S. Richards, R. A. Gibbs, and D. Mittelman. Analysis of microsatellite variation in Drosophila melanogaster with population-scale genome sequencing. *PLoS ONE*, 7(3):1–9, 2012.

[35] A. J. Friedstein, K. V. Petrakova, A. I. Kurolesova, and G. P. Frolova. Heterotopic transplants of bone marrow. *Transplantation*, 6(2):230–247, 1968.

[36] K. Fujimoto, R. D. Beauchamp, and R. H. Whitehead. Identification and isolation of candidate human colonic clonogenic cells based on cell surface integrin expression. *Gastroenterology*, 123(6):1941–8, 2002.

[37] A. Fungtammasan, G. Ananda, S. E. Hile, M. S. W. Su, C. Sun, R. Harris, P. Medvedev, K. Eckert, and K. D. Makova. Accurate typing of short tandem repeats from genome-wide sequencing data and its applications. *Genome Research*, 125(5):736–749, 2015.

[38] N. Gera, M. Hussain, R. C. Wright, and B. M. Rao. Highly stable binding proteins derived from the hyperthermophilic Sso7d scaffold. *Journal of Molecular Biology*, 409(4):601–616, 2011.

[39] F. Gerbe, J. H. Van Es, L. Makrini, B. Brulin, G. Mellitzer, S. Robine, B. Romagnolo, N. F. Shroyer, J. F. Bourgaux, C. Pignodel, H. Clevers, and P. Jay. Distinct ATOH1 and Neurog3 requirements define tuft cells as a new secretory cell type in the intestinal epithelium. *Journal of Cell Biology*, 192(5):767–780, 2011.

[40] L. C. Greaves, S. L. Preston, P. J. Tadrous, R. W. Taylor, M. J. Barron, D. Oukrif, S. J. Leedham, M. Deheragoda, P. Sasieni, M. R. Novelli, J. A. Z. Jankowski, D. M. Turnbull, N. A. Wright, and S. A. C. McDonald. Mitochondrial DNA mutations are established in human colonic stem cells, and mutated clones expand by crypt fission. *Proc Nat Acad Sci USA*, 103(3):714–9, 2006.

[41] A. Gregorieff, D. Pinto, H. Begthel, O. Destrée, M. Kielman, and H. Clevers. Expression pattern of Wnt signaling components in the adult intestine. *Gastroenterology*, 129(2):626–638, 2005.

[42] D. Griffiths, S. Davies, D. Williams, G. Williams, and E. Williams. Demonstration of somatic mutation and colonic crypt clonality by X-linked enzyme histochemistry. *Nature*, 333:461–463, 1988.

[43] E. Guichoux, L. Lagache, S. Wagner, P. Chaumeil, P. Léger, O. Lepais, C. Lepoittevin, T. Malausa, E. Revardel, F. Salin, and R. J. Petit. Current trends in microsatellite genotyping. *Molecular ecology resources*, 11(4):591–611, 2011.

[44] A. Guilmatre, G. Highnam, C. Borel, D. Mittelman, and A. J. Sharp. Rapid multiplexed genotyping of simple tandem repeats using capture and high-throughput sequencing. *Human Mutation*, 34(9):1304–1311, 2013.

[45] M. Gymrek, D. Golan, S. Rosset, and Y. Erlich. lobSTR: A short tandem repeat profiler for personal genomes. *Genome Research*, 22:1154–1162, 2012.

[46] M. J. Hartenstine, M. F. Goodman, and J. Petruska. Base stacking and even/odd behavior of hairpin loops in DNA triplet repeat slippage and expansion with DNA polymerase. *Journal of Biological Chemistry*, 275(24):18382–18390, 2000.

[47] X. Y. Hauge and M. Litt. A study of the origin of 'shadow bands' seen when typing dinucleotide repeat polymorphisms by the PCR. *Human molecular genetics*, 2(4):411–415, 1993.

[48] J. P. Heath. Epithelial cell migration in the intestine. *Cell biology international*, 20(2):139–146, 1996.

[49] G. Highnam, C. Franck, A. Martin, C. Stephens, A. Puthige, and D. Mittelman. Accurate human microsatellite genotypes from high-throughput resequencing data using informed error profiles. *Nucleic Acids Research*, 41(1):1–7, 2013.

[50] A. Humphries, B. Cereser, L. J. Gay, D. S. J. Miller, B. Das, A. Gutteridge, G. Elia, E. Nye, R. Jeffery, R. Poulsom, M. R. Novelli, M. Rodriguez-Justo, S. A. C. McDonald, N. A. Wright, and T. A. Graham. Lineage tracing reveals multipotent stem cells maintain human adenomas and the pattern of clonal expansion in tumor evolution. *Proc Nat Acad Sci USA*, 110(27):E2490–E2499, 2013.

[51] J. Jiricny. The multifaceted mismatch-repair system. *Nat Rev Mol Cell Biol*, 7(5):335–346, 2006.

[52] L. Kaplinski and M. Remm. MultiPLX: Automatic grouping and evaluation of PCR primers. *Methods in Molecular Biology*, 21(8):1701–1702, 2005.

[53] M. Kellett, C. Potten, and D. Rew. A comparison of in vivo cell proliferation measurements in the intestine of mouse and man. *Epithelial Cell Biol.*, 1(4):147–55, 1992.

[54] S. Kozar, E. Morrissey, A. M. Nicholson, M. van der Heijden, H. I. Zecchini, R. Kemp, S. Tavaré, L. Vermeulen, and D. J. Winton. Continuous clonal labeling reveals small numbers of functional stem cells in intestinal crypts and adenomas. *Cell stem cell*, 13(5):626–33, 2013.

[55] F. Kuhnert, C. R. Davis, H.-T. Wang, P. Chu, M. Lee, J. Yuan, R. Nusse, and C. J. Kuo. Essential requirement for Wnt signaling in proliferation of adult small intestine and colon revealed by adenoviral expression of Dickkopf-1. *Proc Nat Acad Sci USA*, 101(1):266–71, 2004.

[56] D. T. Le, J. N. Uram, H. Wang, B. R. Bartlett, H. Kemberling, A. D. Eyring, A. D. Skora, B. S. Luber, N. S. Azad, D. Laheru, B. Biedrzycki, R. C. Donehower, A. Zaheer, G. A. Fisher, T. S. Crocenzi, J. J. Lee, S. M. Duffy, R. M. Goldberg, A. de la Chapelle, M. Koshiji, F. Bhaijee, T. Huebner, R. H. Hruban, L. D. Wood, N. Cuka, D. M. Pardoll, N. Papadopoulos, K. W. Kinzler, S. Zhou, T. C. Cornish, J. M. Taube, R. A. Anders, J. R. Eshleman, B. Vogelstein, and L. A. Diaz. PD-1 Blockade in Tumors with Mismatch-Repair Deficiency. *The New England Journal of Medicine*, 372(26):2509–20, 2015.

[57] M. Lipkin, B. Bell, and P. Sherlock. Cell Proliferation Kinetics in the Gastrointestinal Tract of Man. I. Cell Renewal in Colon and Rectum. *J Clin Invest.*, 42(6):767–776, 1963.

[58] J. Livet, T. A. Weissman, H. Kang, R. W. Draft, J. Lu, R. A. Bennis, J. R. Sanes, and J. W. Lichtman. Transgenic strategies for combinatorial expression of fluorescent proteins in the nervous system. *Nature*, 450(7166):56–62, 2007.

[59] C. Lopez-Garcia, A. M. Klein, B. D. Simons, and D. J. Winton. Intestinal stem cell replacement follows a pattern of neutral drift. *Science*, 330(6005):822–5, 2010.

[60] P. Markoulatos, N. Siafakas, and M. Moncany. Multiplex polymerase chain reaction: A practical approach. *Journal of Clinical Laboratory Analysis*, 16:47–51, 2002.

[61] A. Mauro. Satellite cell of skeletal muscle fibers. *The Journal of biophysical and biochemical cytology*, 9:493–495, 1961.

[62] P. Mazzarello, A. L. Calligaro, and A. Calligaro. Giulio Bizzozero: a pioneer of cell biology. *Nat Rev Mol Cell Biol*, 2(10):776–781, 2001.

[63] F. P. Moss and C. P. Leblond. Satellite cells as the source of nuclei in muscles of growing rats. *The Anatomical record*, 170(4):421–35, 1971.

[64] V. Murray, C. Monchawin, and P. R. England. The determination of the sequences present in the shadow bands of a dinucleotide repeat PCR. *Nucleic acids research*, 21(10):2395–8, 1993.

[65] M. W. Nachman and S. L. Crowell. Estimate of the mutation rate per nucleotide in humans. *Genetics*, 156(1):297–304, 2000.

[66] P. Nicolas, K. M. Kim, D. Shibata, and S. Tavaré. The stem cell population of the human colon crypt: analysis via methylation patterns. *PLoS Comput Biol*, 3(3):364–374, 2007.

[67] M. Novelli, A. Cossu, D. Oukrif, A. Quaglia, S. Lakhani, R. Poulsom, P. Sasieni, P. Carta, M. Contini, A. Pasca, G. Palmieri, W. Bodmer, F. Tanda, and N. Wright. X-inactivation patch size in human female tissue confounds the assessment of tumor clonality. *Proc Nat Acad Sci USA*, 100(6):3311–3314, 2003.

[68] H. Ohno. Intestinal M cells. *Journal of Biochemistry*, 159(2):151–160, 2015.

[69] T. Ohta and M. Kimura. A model of mutation appropriate to estimate the number of electrophoretically detectable alleles in a finite population. *Genetical Research*, 22(2):201–204, 1973.

[70] N. Paweletz. Walther Flemming: pioneer of mitosis research. *Nat Rev Mol Cell Biol*, 2(1):72–75, 2001.

[71] J. Peña-Diaz and J. Jiricny. Mammalian mismatch repair: Error-free or error-prone? *Trends in Biochemical Sciences*, 37(5):206–214, 2012.

[72] Q. Peng, R. Vijaya Satya, M. Lewis, P. Randad, and Y. Wang. Reducing amplification artifacts in high multiplex amplicon sequencing by using molecular barcodes. *BMC genomics*, 16(1):589, 2015.

[73] D. Pinto, A. Gregorieff, H. Begthel, and H. Clevers. Canonical Wnt signals are essential for homeostasis of the intestinal epithelium service. *Genes & development*, 17:1709–1713, 2003.

[74] F. Pompanon, A. Bonin, E. Bellemain, and P. Taberlet. Genotyping errors: causes, consequences and solutions. *Nature Reviews Genetics*, 6(11):847–59, 2005.

[75] B. Ponder, G. Schmidt, M. Wilkinson, M. Wood, M. Monk, and A. Reid. Derivation of mouse intestinal crypts from single progenitor cells. *Nature*, 313(21):689–691, 1985.

[76] C. S. Potten, M. Kellett, S. a. Roberts, D. a. Rew, and G. D. Wilson. Measurement of in vivo proliferation in human colorectal mucosa using bromodeoxyuridine. *Gut*, 33(1):71–78, 1992.

[77] A. E. Powell, Y. Wang, Y. Li, E. J. Poulin, A. L. Means, M. K. Washington, J. N. Higginbotham, A. Juchheim, N. Prasad, S. E. Levy, Y. Guo, Y. Shyr, B. J. Aronow, K. M. Haigis, J. L. Franklin, and R. J. Coffey. The pan-ErbB negative regulator Lrig1 is an intestinal stem cell marker that functions as a tumor suppressor. *Cell*, 149(1):146–58, 2012.

[78] S. C. Pruitt, A. Freeland, and A. Kudla. Cell cycle heterogeneity in the small intestinal crypt and maintenance of genome integrity. *Stem Cells*, 28(7):1250–1259, 2010.

[79] M. Rak, P. Bénit, D. Chrétien, J. Bouchereau, M. Schiff, R. El-Khoury, A. Tzagoloff, and P. Rustin. Mitochondrial cytochrome c oxidase deficiency. *Clinical Science*, 130(6):393–407, 2016.

[80] M. Ramalho-Santos and H. Willenbring. On the Origin of the Term "Stem Cell". *Cell Stem Cell*, 1(1):35–38, 2007.

[81] L. Ritsma, S. I. J. Ellenbroek, A. Zomer, H. J. Snippert, F. J. de Sauvage, B. D. Simons, H. Clevers, and J. van Rheenen. Intestinal crypt homeostasis revealed at single-stem-cell level by in vivo live imaging. *Nature*, 507(7492):362–5, 2014.

[82] L. Ritsma, E. J. A. Steller, S. I. J. Ellenbroek, O. Kranenburg, I. H. M. Borel Rinkes, and J. van Rheenen. Surgical implantation of an abdominal imaging window for intravital microscopy. *Nature Protocols*, 8(3):583–594, 2013.

[83] M. E. Rothenberg, Y. Nusse, T. Kalisky, J. J. Lee, P. Dalerba, F. Scheeren, N. Lobo, S. Kulkarni, S. Sim, D. Qian, P. A. Beachy, P. J. Pasricha, S. R. Quake, and M. F. Clarke. Identification of a cKit+ Colonic Crypt Base Secretory Cell That Supports Lgr5+ Stem Cells in Mice. *Gastroenterology*, 142(5):1195–1205, 2012.

[84] E. Sangiorgi and M. R. Capecchi. Bmi1 is expressed in vivo in intestinal stem cells. *Nat Genet*, 40(7):915–920, 2008.

[85] T. Sato, J. H. van Es, H. J. Snippert, D. E. Stange, R. G. Vries, M. van den Born, N. Barker, N. F. Shroyer, M. van de Wetering, and H. Clevers. Paneth cells constitute the niche for Lgr5 stem cells in intestinal crypts. *Nature*, 469(7330):415–8, 2011.

[86] T. Sato, R. G. Vries, H. J. Snippert, M. van de Wetering, N. Barker, D. E. Stange, J. H. van Es, A. Abo, P. Kujala, P. J. Peters, and H. Clevers. Single Lgr5 stem cells build crypt-villus structures in vitro without a mesenchymal niche. *Nature*, 459(7244):262–265, 2009.

[87] A. G. Schepers, R. Vries, M. van den Born, M. van de Wetering, and H. Clevers. Lgr5 intestinal stem cells have high telomerase activity and randomly segregate their chromosomes. *The EMBO journal*, 30(6):1104–9, 2011.

[88] D. Shinde, Y. Lai, F. Sun, and N. Arnheim. Taq DNA polymerase slippage mutation rates measured by PCR and quasi-likelihood analysis: (CA/GT)n and (A/T)n microsatellites. *Nucleic Acids Research*, 31(3):974–980, 2003.

[89] H. J. Snippert, L. G. van der Flier, T. Sato, J. H. van Es, M. van den Born, C. Kroon-Veenboer, N. Barker, A. M. Klein, J. van Rheenen, B. D. Simons, and H. Clevers. Intestinal crypt homeostasis results from neutral competition between symmetrically dividing Lgr5 stem cells. *Cell*, 143(1):134–144, 2010.

[90] A. Ståhlberg, P. M. Krzyzanowski, J. B. Jackson, M. Egyud, L. Stein, and T. E. Godfrey. Simple, multiplexed, PCR-based barcoding of DNA enables sensitive mutation detection in liquid biopsies using sequencing. *Nucleic Acids Research*, 1(1):1–7, 2016.

[91] J. C. Strafford. Genetic testing for lynch syndrome, an inherited cancer of the bowel, endometrium, and ovary. *Reviews in obstetrics & gynecology*, 5(1):42–9, 2012.

[92] M. Strand, T. a. Prolla, R. M. Liskay, and T. D. Petes. Destabilization of tracts of simple repetitive DNA in yeast by mutations affecting DNA mismatch repair. *Nature*, 365(6443):274–276, 1993.

[93] S. Subramanian, R. K. Mishra, and L. Singh. Genome-wide analysis of microsatellite repeats in humans: their abundance and density in specific genomic regions. *Genome biology*, 4(2):R13, 2003.

[94] F. Syed, H. Grunenwald, and N. Caruccio. Next-generation sequencing library preparation: Simultaneous fragmentation and tagging using in vitro transposition. *Nature Methods*, 6(11):i–ii, 2009.

[95] N. Takeda, R. Jain, M. R. LeBoeuf, Q. Wang, M. M. Lu, and J. a. Epstein. Interconversion between intestinal stem cell populations in distinct niches. *Science*, 334(6061):1420–4, 2011.

[96] R. W. Taylor, M. J. Barron, G. M. Borthwick, A. Gospel, P. F. Chinnery, D. C. Samuels, G. a. Taylor, S. M. Plusa, S. J. Needham, L. C. Greaves, T. B. L. Kirkwood, and D. M. Turnbull. Mitochondrial DNA mutations in human colonic crypt stem cells. *J Clin Invest*, 112(9):1351–1360, 2003.

[97] P. Tetteh, O. Basak, H. Farin, K. Wiebrands, K. Kretzschmar, H. Begthel, M. van den Born, J. Korving, F. de Sauvage, J. van Es, A. van Oudenaarden, and H. Clevers. Replacement of Lost Lgr5-Positive Stem Cells through Plasticity of Their Enterocyte-Lineage Daughters. *Cell Stem Cell*, 18(2):203–213, 2016.

[98] E. D. Thomas, H. L. Lochte, W. C. Lu, and J. W. Ferrebee. Intravenous infusion of bone marrow in patients receiving radiation and chemotherap. *The New England journal of medicine*, 257:491–496, 1957.

[99] M. Thompson, K. A. Fleming, D. J. Evans, R. Fundele, M. A. Surani, and N. A. Wright. Gastric endocrine cells share a clonal origin with other gut cell lineages. *Development*, 110:477–481, 1990.

[100] J. E. Till and E. A. McCulloch. A Direct Measurement of the Radiation Sensitivity of Normal Mouse Bone Marrow Cells. *Radiation Research*, 14(2):213–222, 1961.

[101] J. E. Till, E. A. McCulloch, and L. Siminovitch. A Stochastic Model of Stem Cell Proliferation, Based on the Growth of Spleen Colony-Forming Cells. *Proc Nat Acad Sci USA*, 51(1):29–36, 1964.

[102] L. G. van der Flier and H. Clevers. Stem cells, self-renewal, and differentiation in the intestinal epithelium. *Annual review of physiology*, 71:241–60, 2009.

[103] J. H. van Es, M. E. van Gijn, O. Riccio, M. van den Born, M. Vooijs, H. Begthel, M. Cozijnsen, S. Robine, D. J. Winton, F. Radtke, and H. C. Clevers. Notch/gamma-secretase inhibition turns proliferative cells in intestinal crypts and adenomas into goblet cells. *Nature*, 435(7044):959–63, 2005.

[104] D. J. Winton, M. A. Blount, and B. A. Ponder. A clonal marker induced by mutation in mouse intestinal epithelium. *Nature*, 333(6172):463–466, 1988.

[105] D. J. Winton and B. A. Ponder. Stem-cell organization in mouse small intestine. *Proc Biol Sci*, 241(1300):13–18, 1990.

[106] V. W. Y. Wong, D. E. Stange, M. E. Page, S. Buczacki, A. Wabik, S. Itami, M. van de Wetering, R. Poulsom, N. A. Wright, M. W. B. Trotter, F. M. Watt, D. J. Winton, H. Clevers, and K. B. Jensen. Lrig1 controls intestinal stem-cell homeostasis by negative regulation of ErbB signalling. *Nature cell biology*, 14(4):401–8, 2012.

[107] N. A. Yamada, G. A. Smith, A. Castro, C. N. Roques, J. C. Boyer, and R. A. Farber. Relative rates of insertion and deletion mutations in dinucleotide repeats of various lengths in mismatch repair proficient mouse and mismatch repair deficient human cells. *Mutat Res*, 499(2):213–225, 2002.

[108] K. S. Yan, L. a. Chia, X. Li, A. Ootani, J. Su, J. Y. Lee, N. Su, Y. Luo, S. C. Heilshorn, M. R. Amieva, E. Sangiorgi, M. R. Capecchi, and C. J. Kuo. The intestinal stem cell markers Bmi1 and Lgr5 identify two functionally distinct populations. *Proc Nat Acad Sci USA*, 109(2):466–71, 2012.

[109] Y. Yatabe, S. Tavare, and D. Shibata. Investigating stem cells in human colon by using methylation patterns. *Proc Nat Acad Sci USA*, 98(19):10839–10844, 2001.

[110] F. M. You, N. Huo, Y. Q. Gu, M.-C. Luo, Y. Ma, D. Hane, G. R. Lazo, J. Dvorak, and O. D. Anderson. BatchPrimer3: a high throughput web application for PCR and sequencing primer design. *BMC bioinformatics*, 9:253, 2008.

[111] M. Zavodna, A. Bagshaw, R. Brauning, and N. J. Gemmell. The accuracy, feasibility and challenges of sequencing short tandem repeats using next-generation sequencing platforms. *PLoS ONE*, 9(12):1–14, 2014.

Appendix A

Primer sequences for mouse and human microsatellite analysis

A.1 Next Generation Sequencing adaptors

Table A.1 gives the sequences of adaptors added to the 5'-end of the locus specific primers: when a CS1/CS2 adaptor is present, the locus specific primer is given the prefix 'ADAP_' whilst when the M13 adaptor is present, the locus specific primer is given the prefix 'M13_'. Table A.2 gives the primer sequences of the custom made M13 indexing primers: i5 primers are coupled with i7 primers to form a primer pair. By matching i5 primers with different i7 primers, it is possible to generate 384 unique indexing sequences.

Adaptor	**Forward primer**	**Reverse primer**
CS1/CS2	ACACTGACGACATGGTTCTACA	TACGGTAGCAGAGACTTGGTCT
M13	TGTAAAACGACGGCCAGT	CAGGAAACAGCTATGACC

Table A.1 Sequence of Next Generation Sequencing adaptors added to the 5'-end of the forward and reverse amplification primers as indicated.

Index name	Index sequence	Primer
M13_i5_S502	CTCTCTAT	AATGATACGGCGACCACCGAGATCTACACCTCTCTATTCGTCGGCAGCGTCAGATGTGTATAAGAGACAGtgtaaaacgacggccagt
M13_i5_S503	TATCCTCT	AATGATACGGCGACCACCGAGATCTACACTATCCTCTTCGTCGGCAGCGTCAGATGTGTATAAGAGACAGtgtaaaacgacggccagt
M13_i5_S505	GTAAGGAG	AATGATACGGCGACCACCGAGATCTACACGTAAGGAGTCGTCGGCAGCGTCAGATGTGTATAAGAGACAGtgtaaaacgacggccagt
M13_i5_S506	ACTGCATA	AATGATACGGCGACCACCGAGATCTACACACTGCATATCGTCGGCAGCGTCAGATGTGTATAAGAGACAGtgtaaaacgacggccagt
M13_i5_S507	AAGGAGTA	AATGATACGGCGACCACCGAGATCTACACAAGGAGTATCGTCGGCAGCGTCAGATGTGTATAAGAGACAGtgtaaaacgacggccagt
M13_i5_S508	CTAAGCCT	AATGATACGGCGACCACCGAGATCTACACCTAAGCCTTCGTCGGCAGCGTCAGATGTGTATAAGAGACAGtgtaaaacgacggccagt
M13_i5_S510	CGTCTAAT	AATGATACGGCGACCACCGAGATCTACACCGTCTAATTCGTCGGCAGCGTCAGATGTGTATAAGAGACAGtgtaaaacgacggccagt
M13_i5_S511	TCTCTCCG	AATGATACGGCGACCACCGAGATCTACACTCTCTCCGTCGTCGGCAGCGTCAGATGTGTATAAGAGACAGtgtaaaacgacggccagt
M13_i5_S513	TCGACTAG	AATGATACGGCGACCACCGAGATCTACACTCGACTAGTCGTCGGCAGCGTCAGATGTGTATAAGAGACAGtgtaaaacgacggccagt
M13_i5_S515	TTCTAGCT	AATGATACGGCGACCACCGAGATCTACACTTCTAGCTTCGTCGGCAGCGTCAGATGTGTATAAGAGACAGtgtaaaacgacggccagt
M13_i5_S516	CCTAGAGT	AATGATACGGCGACCACCGAGATCTACACCCTAGAGTTCGTCGGCAGCGTCAGATGTGTATAAGAGACAGtgtaaaacgacggccagt
M13_i5_S517	GCGTAAGA	AATGATACGGCGACCACCGAGATCTACACGCGTAAGATCGTCGGCAGCGTCAGATGTGTATAAGAGACAGtgtaaaacgacggccagt
M13_i5_S518	CTATTAAG	AATGATACGGCGACCACCGAGATCTACACCTATTAAGTCGTCGGCAGCGTCAGATGTGTATAAGAGACAGtgtaaaacgacggccagt
M13_i5_S520	AAGGCTAT	AATGATACGGCGACCACCGAGATCTACACAAGGCTATTCGTCGGCAGCGTCAGATGTGTATAAGAGACAGtgtaaaacgacggccagt
M13_i5_S521	GAGCCTTA	AATGATACGGCGACCACCGAGATCTACACGAGCCTTATCGTCGGCAGCGTCAGATGTGTATAAGAGACAGtgtaaaacgacggccagt
M13_i5_S522	TTATGCGA	AATGATACGGCGACCACCGAGATCTACACTTATGCGATCGTCGGCAGCGTCAGATGTGTATAAGAGACAGtgtaaaacgacggccagt
M13_i7_N701	TCGCCTTA	CAAGCAGAAGACGGCATACGAGATTCGCCTTAGTCTCGTGGGCTCGGAGATGTGTATAAGAGACAGcaggaaacagctatgacc
M13_i7_N702	CTAGTACG	CAAGCAGAAGACGGCATACGAGATCTAGTACGGTCTCGTGGGCTCGGAGATGTGTATAAGAGACAGcaggaaacagctatgacc

M13_i7_N703	TTCTGCCT	CAAGCAGAAGACGGCATACGAGATTTCTGCCTGTCTCGTGGGCTCGGAGATGTGTATAAGAGACAGcaggaaacagctatgacc
M13_i7_N704	GCTCAGGA	CAAGCAGAAGACGGCATACGAGATGCTCAGGAGTCTCGTGGGCTCGGAGATGTGTATAAGAGACAGcaggaaacagctatgacc
M13_i7_N705	AGGAGTCC	CAAGCAGAAGACGGCATACGAGATAGGAGTCCGTCTCGTGGGCTCGGAGATGTGTATAAGAGACAGcaggaaacagctatgacc
M13_i7_N706	CATGCCTA	CAAGCAGAAGACGGCATACGAGATCATGCCTAGTCTCGTGGGCTCGGAGATGTGTATAAGAGACAGcaggaaacagctatgacc
M13_i7_N707	GTAGAGAG	CAAGCAGAAGACGGCATACGAGATGTAGAGAGGTCTCGTGGGCTCGGAGATGTGTATAAGAGACAGcaggaaacagctatgacc
M13_i7_N710	CAGCCTCG	CAAGCAGAAGACGGCATACGAGATCAGCCTCGGTCTCGTGGGCTCGGAGATGTGTATAAGAGACAGcaggaaacagctatgacc
M13_i7_N711	TGCCTCTT	CAAGCAGAAGACGGCATACGAGATTGCCTCTTGTCTCGTGGGCTCGGAGATGTGTATAAGAGACAGcaggaaacagctatgacc
M13_i7_N712	TCCTCTAC	CAAGCAGAAGACGGCATACGAGATTCCTCTACGTCTCGTGGGCTCGGAGATGTGTATAAGAGACAGcaggaaacagctatgacc
M13_i7_N714	TCATGAGC	CAAGCAGAAGACGGCATACGAGATTCATGAGCGTCTCGTGGGCTCGGAGATGTGTATAAGAGACAGcaggaaacagctatgacc
M13_i7_N715	CCTGAGAT	CAAGCAGAAGACGGCATACGAGATCCTGAGATGTCTCGTGGGCTCGGAGATGTGTATAAGAGACAGcaggaaacagctatgacc
M13_i7_N716	TAGCGAGT	CAAGCAGAAGACGGCATACGAGATTAGCGAGTGTCTCGTGGGCTCGGAGATGTGTATAAGAGACAGcaggaaacagctatgacc
M13_i7_N718	GTAGCTCC	CAAGCAGAAGACGGCATACGAGATGTAGCTCCGTCTCGTGGGCTCGGAGATGTGTATAAGAGACAGcaggaaacagctatgacc
M13_i7_N719	TACTACGC	CAAGCAGAAGACGGCATACGAGATTACTACGCGTCTCGTGGGCTCGGAGATGTGTATAAGAGACAGcaggaaacagctatgacc
M13_i7_N720	AGGCTCCG	CAAGCAGAAGACGGCATACGAGATAGGCTCCGGTCTCGTGGGCTCGGAGATGTGTATAAGAGACAGcaggaaacagctatgacc
M13_i7_N721	GCAGCGTA	CAAGCAGAAGACGGCATACGAGATGCAGCGTAGTCTCGTGGGCTCGGAGATGTGTATAAGAGACAGcaggaaacagctatgacc
M13_i7_N722	CTGCGCAT	CAAGCAGAAGACGGCATACGAGATCTGCGCATGTCTCGTGGGCTCGGAGATGTGTATAAGAGACAGcaggaaacagctatgacc
M13_i7_N723	GAGCGCTA	CAAGCAGAAGACGGCATACGAGATGAGCGCTAGTCTCGTGGGCTCGGAGATGTGTATAAGAGACAGcaggaaacagctatgacc
M13_i7_N724	CGCTCAGT	CAAGCAGAAGACGGCATACGAGATCGCTCAGTGTCTCGTGGGCTCGGAGATGTGTATAAGAGACAGcaggaaacagctatgacc
M13_i7_N726	GTCTTAGG	CAAGCAGAAGACGGCATACGAGATGTCTTAGGGTCTCGTGGGCTCGGAGATGTGTATAAGAGACAGcaggaaacagctatgacc
M13_i7_N727	ACTGATCG	CAAGCAGAAGACGGCATACGAGATACTGATCGGTCTCGTGGGCTCGGAGATGTGTATAAGAGACAGcaggaaacagctatgacc

M13_i7_N728	TAGCTGCA	CAAGCAGAAGACGGCATACGAGATTAGCTGCAGTCTCGTGGGCTCGGAGATGTGTATAAGAGACAGcaggaaacagctatgacc
M13_i7_N729	GACGTCGA	CAAGCAGAAGACGGCATACGAGATGACGTCGAGTCTCGTGGGCTCGGAGATGTGTATAAGAGACAGcaggaaacagctatgacc

Table A.2 Sequences for indexing primers used for M13 indexing. Lower case lettering represents regions complimentary to the M13 adaptor.

A.2 Mouse primers

All of the primers listed here were part of the final multiplex PCR group named 'M13_33' or 'M13_33_YFP' when primers for the transgenic $[CA]_{30}$ were included. Each primer is written in the conventional 5' to 3' direction. Sequencing adaptors are added at the 5' end.

Loci	Forward primer	Reverse primer	Final conc. (μM)	Amplicon length (bp)
a2_2742	TCTTGCTCTTTCCTACACTGTT	CTCTGTGTTCCCAGCCCTAG	0.18	151
s16_4295	TTTGCATACTCAGGAAACCCA	TCACAGACACTAGAGCAGCA	0.34	160
a4_1365	AAATCTCTGAACTTCCTCCCTG	AGCTGTCACGTGATTCCAGA	0.26	152
s2_1206	GTAGTCAACCGAAATACACGAAA	ACAAGATCAGCAAGCTCTCAAC	0.28	166
s8_3334	AGATGTCTTCTTCCAACCTCCA	GGTCTGTTTGATTTCATGATGCC	0.16	156
s9_8328	GGACCATCACCCTAAGAAACAC	GACACTGCCCCTTACACTGT	0.17	170
a6_42765	GGGTTGTTGGTCAGGTTTAAGGTAGCA	CCCAAGACTTCATATATAACTGAGGC	0.09	184
sX_13950	TGGCCAATCCAAAAATACAGACACAG	TGTTTGGGCACGAACCTTTATAAGCTG	0.23	170
a10_2051	GCATCTGTTGGAGAATGAATATGTTGT	ACATCTGAGCCATCTTTGCAAGTCTGT	0.27	196
s8_6742	GAGGAGCTAGCTACACTCGG	ACGGTGAGTTGGGTGTCTTT	0.43	174
s15_7506	TGTCCTCTGAACTTTGTATGCAG	GCCCTTTCTTCTGTGCCAC	0.14	180
s8_27334	GCATCGATTTTCATGGGGATTTTGATA	AAGTTTCCACTTCAATTTGCCTTCCTG	0.28	170
s9_69438	TAGTCTAATCAGAGAGCCCAGGCCAGT	CAGGGAGAGCCAATGAATGAGTCAGTA	0.20	185
s9_4554	AGCTGTCAGTTTTACCCATGGTCATCA	ATGGGCATACAGACAAACCAGTAGGTG	0.15	177
s1_EYFP	TGTCATACTTATCCTGTCCCTTTTT	AACAGCTCCTCGCCCTTG	0.57	176

Table A.3 Primer sequences for primer pairs contained within mouse multiplex group M13_33 and M13_33_YFP (when the group includes the primers for the transgenic $[CA]_{30}$).

A.3 Human primers

All of the primers listed here were part of the final multiplex PCR group named 'hsM13_53'. Each primer is written in the conventional 5' to 3' direction. Sequencing adaptors are added at the 5' end.

Loci	Forward primer	Reverse primer	Final conc. (μM)	Amplicon length (bp)
s3_50602	AAAGGAATGTTGGCCAAGTG	AAAGGAATGTTGGCCAAGTG	0.09	180
s9_39733	GAAATTTCATGCCAATTTTAATACA	GAGTCCAGCACTGTCTGCAA	0.47	181
s12_1242	CAACCAGAGGTGCGGATACT	AGACCCATTTCCTCCAATCC	0.32	181
s14_6560	TTCCATCTAGGATGCTGCAA	TTGCAAAAATGTGGAACCAA	0.47	180
s15_8821	TCTCCCAGCTCCTCCTACCT	CAGCCACAGAGTGGGACATA	0.05	169
s17_6742	CCTGCACCCAGGTGAAATAA	GGGGTTCATTCTCCCATACC	0.32	180
a8_86853	TGCCCATGGGTAAATGAAAC	TCCCAAGGGCACACTGTT	0.08	154
a3_12858	TACACCCTGGGAACGAGTTT	TGGCAGGTGGTGATAATGAA	0.16	161
a9_99288	GGAGTAAAAGAAAGCCACCAAA	AGCTGGGGTCACATCAGAAC	0.47	178
a7_29069	TCGGCAATTCATTGCTATCC	AAAGAGTAGTAACTCAGAGCTGGAGA	0.08	183
a17_6874	CATTCCTTGGTTGCATGACT	TCCTGGGTCTAGCATAAGTCTTAAA	0.32	185
s22_4402	CAGGATGGGCTCTAATCCAA	TGTTCCTCGGCCTGTAGAAG	0.16	163
s14_1937	CCAATGATGAAACTGGCTCA	GGAGGGTCTCTCTCCACTTTC	0.16	181
s2_87089	ATCGTTTCACATTTGAACTCTTT	CTGCAACGGTCCCCTTCT	0.32	176
s13_7008	AGCAGTGCAAGAATGGAATAA	CCCAGCCTGGACCTTAAATC	0.47	191
s12_6308	TTTCCCTCATGGCTACTCTTG	AACTGCAGTGGCAAGAAACC	0.08	177
a3_46762	GGTTTGTAACACGTATGCAAGAA	GAGCGAAACTCTGTCTCAAAAA	0.47	197
s14_6792	AAAGCAAGGACATAACCCACA	GAGGACCTGCTGTAATTCCCTA	0.32	169
a9_13528	TCAACTGCCAGCAAAAGTTAAA	CAAAACACGTAGAGGGTGGAA	0.32	177
a4_10696	AAAGGTGCCAAGAACATACATT	TTTTTGTAGATGGTGAGAGATGG	0.32	175
s17_1170	CTGTCCTCAGAGGCCTTCC	GAGGCTCTGTCAGGGTTACCT	0.05	172

Table A.4 Primer sequences for primer pairs contained within human multiplex group hsM13_53.

Appendix B

Genomic information for mouse and human microsatellites used for clone size analysis

All genomic loci were identified, using the custom Perl script described in Section 2.10.1, in the mm9 build of the mouse reference genome and hg38 build of the human reference genome. Each loci included for analysis in the final multiplex groups were searched for using the UCSC Genome Browser. The UCSC Genes genome annotation was used to identify murine coding regions whilst the RefSeq genome annotation was used to identify human coding regions.

B.1 Mouse microsatellite loci

Table B.1 gives the details of all loci analysed using the M13_33 multiplex group.

B.2 Human microsatellite loci

Table B.2 gives the details of all loci analysed using the hsM13_53 multiplex group.

Chromosome	Start	End	Type	Distance to nearest gene (nucleotides)	Nearest gene
2	27425428	274254988	Intergenic	23820	BC050122
2	120629581	120629641	4th intron	0	Ttbk2
4	136517568	136517628	Intergenic	30605	Epha8
6	42764938	42764998	Intergenic	2534	Olfr450
8	27334239	27334299	Intergenic	60968	Thap1
8	33339937	33339997	Intergenic	77570	AK143160
8	67422899	67422959	Intergenic	48307	Tmem192
9	69438634	69438694	Intergenic	99322	Anxa2
9	83283058	83283118	Intergenic	5664	Lca5
10	20517849	20517909	Intergenic	73347	Pde7b
15	75067328	75067388	Intergenic	20213	BC025446
16	42959516	42959576	2nd intron	0	Zbtb20
19	45540103	45540163	4th intron	0	Btrc
X	139502950	139503010	Intergenic	31136	Rgag1

Table B.1 Genomic information for all loci analysed within the M13_33 multiplex group.

Chromosome	Start	End	Type	Distance to nearest gene (nucleotides)	Nearest gene
2	87089769	87089829	Intergenic	67986	PLGLB1
3	46762346	46762406	Intergenic	44522	PRSS50
3	50602054	50602114	Intergenic	4442	CISH
3	128580366	128580426	Intergenic	39985	RPN1
4	106962493	106962553	1st intron	0	DKK2
7	29069768	29069828	9th intron	0	CPVL
8	86853133	86853193	Intergenic	13303	CNBD1
9	13528207	13528267	Intergenic	248643	MPDZ
9	39733059	39733119	Intergenic	80665	KGFLP2
9	99288207	99288267	Intergenic	66313	ALG2
12	6308773	6308833	Intergenic	3394	PLEKHG6
12	124203894	124203954	3rd intron	0	FAM101A
13	70082309	70082369	1st intron	0	KLHL1
14	19378602	19378662	Intergenic	0	POTEM
14	65602343	65602403	3rd intron	0	FUT8
14	67921563	67921623	7th intron	0	RAD51B
15	88212700	88212760	3rd intron	0	NTRK3
17	11707955	11708015	26th intron	0	DNAH9
17	67425203	67425263	1st intron	0	PITPNC1
17	68740442	68740502	Intergenic	127643	ABCA8
22	44021757	44021817	2nd intron	0	PARVB

Table B.2 Genomic information for all loci analysed within the hsM13_53 multiplex group.